Int J Ethics Soc 2025; 6(4): 1-10

doi 10.22034/ijethics.6.4.1

INTERNATIONAL JOURNAL OF
Ethics &
Society

Review Article

Ethical Analysis of the Responsibility Gap in Artificial Intelligence

Eva Schur*, Anna Brouns, Peter Lee

Department of Artificial Intelligence and Cybersecurity, Faculty of Technical Sciences, University of Klagenfurt, Austria.

Corresponding Author: Eva Schur, Department of Artificial Intelligence and Cybersecurity, Faculty of Technical Sciences, University of Klagenfurt, Austria. E-mail: evaschur10@gmail.com

Received 09 Oct 2024 **Accepted** 25 Nov 2024 **Online Published** 24 Jan 2025

Abstract

Introduction: The concept of the "responsibility gap" in artificial intelligence (AI) was first raised in philosophical discussions to reflect concerns that learning and partially autonomous technologies may make it more difficult or impossible to attribute moral blame to individuals for adverse events. This is because in addition to designers, the environment and users also participate in the development process. This ambiguity and complexity sometimes make it seem that the output of these technologies is beyond the control of human individuals and that no one can be held responsible for it, which is known as the "responsibility gap". In this article, the issue of the responsibility gap in artificial intelligence technologies will be explained and strategies for the responsible development of artificial intelligence that prevent such a gap from occurring as much as possible are presented.

Material and Methods: The present article examined responsibility gap in AI. In order to achieve this goal, related articles and books were examined.

Conclusion: There have been various responses to the issue of the responsibility gap. Some believe that society can hold the technology responsible for its outcomes. Others disagree. Accordingly, only the human actors involved in the development of these technologies can be held responsible, and they should be expected to use their freedom and awareness to shape the path of technological development in a way that prevents undesirable and unethical events. In summary, the three principles of routing, tracking, and engaging public opinion and attention to public emotions in policymaking can be useful as three effective strategies for the responsible development of AI technologies.

Keywords: *Ethics, Responsibility Gap, Artificial Intelligence*

How to Cite: Schur E, Brouns A, Lee P. Ethical analysis of the responsibility gap in artificial intelligence, Int J Ethics Soc. 2025;6(4): 1-10. doi: 10.22034/ijethics.6.4.1

INTRODUCTION

The increasing advancement of technology and its rapid spread in the world have made the issue of responsibility and technology an important issue for thinkers. The central question is that, despite the complex processes of design, development and use of technology, who or what institutions are responsible for the consequences of the applications of technology? This issue still exists and is the subject of much debate regarding established and old technologies such as chemical factories, automobiles and airplanes, and medicines, whose processes are more or less known and transparent. But the issue becomes more complicated when we deal with emerging and complex technologies such as artificial intelligence.

In addition to the traditional complexities of socio-technical networks, these technologies have additional complexities in their development processes. The new generation of artificial intelligence is trained by big data, which means that many environmental factors that may be out of control affect the performance of these technologies. In addition, the power of AI in data analysis, decision-making, and control of large systems makes us entrust important roles to AI-based systems; roles such as: managing transportation systems, managing public health systems, programming, or perhaps in the not-too-distant future, scientific theorizing. Our previous question about how to distribute responsibility for the consequences of the application of technology has now taken on a more complex and frightening face. It seems that humanity is on the verge of a huge transformation, and the helm of guiding important parts of human civilization is being entrusted to AI technologies that have many dark and unknown dimensions. Therefore, the extent of this delegation and the measures that must be taken to use these technologies are of great importance. In this article, the issue of the responsibility gap regarding AI technologies will be explained and strategies for the responsible development of AI that will prevent such a gap from occurring as much as possible are presented.

MATERIAL AND METHODS

The present article examined responsibility gap in AI through a systematic review of related articles. In order to achieve this goal, related articles and books were examined.

DISCUSSION

What is the responsibility gap?

Responsibility is a fundamental concept in ethics. It establishes a connection between the state of the external world and the agents who have the power to influence it, and it is on this basis that moral action, punishment, and reward make sense. Different analyses of responsibility have been presented by philosophers of different eras. Responsibility for a situation sometimes means that a person can be blamed for blamability, accountability and culpability, if the situation is positive and desirable, praised and encouraged for it. For these reactions to a person's action to be reasonable, it is necessary that the situation in question is in some way the product of the person's actions. According to Aristotle, we can consider a person responsible for an action (and its consequences) in circumstances where he has sufficient knowledge and control over that action. That is, he must have both done the act intentionally and with awareness of the consequences, and he must have had the power to do something else [1, 2]. Others have defined the idea of responsibility using similar concepts.

In a world where various technologies play a mediating role in almost all human actions, the issue of responsibility has taken on a new dimension. Suppose you witness a car crash with a pedestrian. The car hits a pedestrian on the crosswalk while the light is red. Your initial judgment is probably that the driver of the car is at fault for not braking in time. But if you know that this car is the product of a factory whose entire production has minor defects in the brakes and brakes with a half-second delay, your opinion about the driver's responsibility may change. Also, if you know that there is a minor problem with the traffic lights and the pedestrian light turns green before the car light turns red, or if you see that the pedestrian line on the ground has faded and is not easy for drivers to recognize, you may still reconsider the level of responsibility you attribute to the driver. The car factory, the traffic lights, and the crosswalk are all technical dimensions of the incident that play a role in how responsibility is assigned. They can affect the degree of agency and power of the user to prevent the accident. As the parameters involved in performing an action increase, we can attribute an action to a human agent with less certainty.

This uncertainty increases significantly with artificial intelligence technologies.

Some researchers point to numerous examples of AI technologies in which the programmer and the user play a very minimal role in the final action and say that technological progress has led to a new phenomenon that they call the accountability gap. This gap, which deepens as technological networks become more numerous and more complex, is the product of the reduction of control by human agents [3]. As a result of this phenomenon, we are faced with incidents for which no one can be held responsible or forced to compensate for the damage.

New generations of AI, which are learning machines, can decide and act on some tasks without human intervention. They obey rules that are not determined during their design and production process, but are formed by the machine itself and based on environmental feedback when the machine starts to act. The flexibility of the machine to new conditions dictates that the adaptation process should take place in it. That is, it can move away from its initial settings and planning. This adaptation and learning is possible only through trial and error. The question is, who is to be held responsible for these (sometimes) unavoidable errors? Given the fundamental limitations that the user and the programmer face in their awareness of the machine's operation and control of its outputs, they cannot be held fully responsible for the error. Because neither has sufficient control over the machine's actions [3]. Artificial intelligence systems interact with objects, humans, and social identities such as laws and institutions, and through these interactions gain insight into and change the environment. Unlike simpler processors (e.g., desktop computers), they have direct access to sensory stimuli and their symbolic representations, and they operate in the same environment as humans [3].

Mathias sees the role of human agents in these technologies as different from their role in other technologies. Here, the programmer no longer directly loads instructions into the software so that the computer follows his algorithms line by line, but rather develops an autonomous AI system. In earlier versions of AI, there was more transparency and we could examine the code line by line, explain it, and find its errors, but learning AI models have a different way of working. In the various approaches to AI programming (from symbolic systems, connectionist systems and neural networks, genetic algorithms, and finally autonomous agents), the programmer's control over the program's performance and final output gradually decreases. They have a program that consists of a set of axioms and inference rules, and it is the AI itself that is involved in the process of drawing conclusions and acting. What happens and how exactly the learning and decision-making process will proceed is to some extent beyond the programmer's prediction and knowledge [3]. Although AI systems are still under partial human supervision and have limited autonomy, their level of autonomy is increasing day by day, they are being assigned more complex tasks and are being used in contexts where there are other unpredictable factors and consequences and the possibility of unpredictable risks. The multiplicity and complexity of the environment and people who interact with AI makes the consequences of the operation of this technology neither predictable nor legally traceable [3]. We have an increasing number of situations in which the possibility of a human agent supervising a working machine is ruled out, either in principle or for economic reasons; For example, when the machine has an information advantage over the user (navigation computers in vehicles that have satellite information), or when their processing speed is very high and there are multiple operational factors (e.g., analyzing the movement of several elevators in a tower of several dozen floors) that are beyond the control of human agents.

Consequently, the option of eliminating such technologies seems to be unavailable to us. We can neither demand these functions from simpler systems nor forego their benefits. Matthias recommends that in order to avoid unjustly holding humans responsible for actions over which they have no control, we must find a way to bridge this responsibility gap in moral and legal practice [3]. In the next section, we will examine the various responses that philosophers have given to the problem of the responsibility gap.

Can AI be held accountable?

In this section, we will review some of the proposals that have been made to resolve the issue of the responsibility gap. One of the proposals to resolve this issue is to attribute responsibility and accountability for the actions and harms caused by AI to itself, which requires that AI fulfill the conditions of responsibility - which were previously raised - so that we include and accept AI systems as members of our moral communities and as moral agents.

Various answers have been proposed regarding the question of how and under what conditions an entity can be considered a moral agent or be held accountable: According to the first view, which we call the "property focused ", an entity is considered a moral agent if it has certain features. Properties such as consciousness [4], autonomy [5], specific intentional states [6], practical rationality [7], specific emotional states, higher-order intentionality [8], or a combination of these [9] can be considered properties of a moral agent. The second view, however, which we call the "relative approach," says that the moral status of an entity depends on how we perceive it, how we interact with it, and how we respond to it in specific encounters with it. Thus, it is the way we engage with it that determines its moral agency.

Artificial intelligence is rapidly moving towards situations in which the agent, rather than an AI, can be considered a moral situation. For example, a nurse has certain moral responsibilities and rights when she treats her patients. Do these rights and responsibilities also apply if the AI is a caregiver? In this section, we will analyze and enumerate the answers to the question, "Can artificial intelligence be given agency and responsibility, thereby solving the responsibility gap?" from the perspectives of some philosophers and technologists. In general, the answers can be divided into four categories:

- ***Group 1: AI can be responsible and agentive***

If we think of technology as the manipulation of nature to achieve human goals, we can consider guide dogs as a type of technology. This technology is inherently intelligent and possibly has some kind of consciousness. Unlike the user-tool model, we are dealing here with a complex relationship between the trainer, the guide dog, and the blind person the dog is supposed to help. Most of us would consider the dog's assistance to be morally valuable. But what exactly is this moral admiration attributed to? Some researchers believe that both the trainer and the dog share in it. We admire the skill and dedication of the trainers, and we admire the work of the dog. We describe the relationship between humans and dogs in the same terms that we describe the relationship between two humans. Beyond the work of the trainers, the work of the dog itself has moral value [9].

According to these researchers, although no robot has the cognitive power of a dog, if we consider the group of technologies as a spectrum, today's robots are more like a dog than simple tools such as a hammer. So it is not far from the mind to think about their value and moral agency. Autonomous artificial intelligence (in the same conventional sense used by engineers), which can make some decisions by its own programming, has moral agency. If the actions of an entity are interactive, adaptable to the surrounding environment and to some extent independent of it, it is sufficient for this entity to have agency [9]. In summary, the three conditions that such researchers propose for the moral agency of artificial intelligence are as follows: 1) the artificial intelligence must be sufficiently autonomous in relation to the programmer or machine operator. 2) its behavior must be such that in order to analyze it we must attribute some kind of moral or immoral intention to it. and 3) perform social roles that involve some social responsibilities, and the only way we can give meaning to his work is to attribute to him a sense of responsibility. According to Salinas, by fulfilling all these conditions, AI has both rights and moral responsibilities, regardless of whether it is a person or not [9, 10].

- ***Group 2: AI in the future can have responsibility and agency***

Some researchers believe that AI in the future can have agency and responsibility. They say that a machine can have malicious intent or a sinful mental state that includes motivational states such as purpose, cognitive states such as belief, or non-mental states such as fault and negligence [11]. But in order for AI to be morally punishable, this technology must also have higher-order intentionalities. That is, it must have beliefs about beliefs and desires about its desires; beliefs about its fears, thoughts, hopes, etc. Dennett does not believe that we have such machines today. But he does not deny that such machines will not exist in the future either.

- ***Group Three: Responsibility and agency cannot be delegated to AI***

According to the instrumentalist view in the philosophy of technology, the responsibility for the current consequences caused by a technology lies with its user. According to this view, just as a person can be killed with a knife, a person can also be saved from certain death with surgery. The technologies that are intertwined in the fabric of our lives are not as simple and tangible as a knife. In this model, if the use of technology leads to harm, we usually blame the user and not the tool. According to this line of reasoning, the robot is no longer a moral agent and is at best a tool that advances the moral interests of others.

Therefore, some researchers do not believe in assigning responsibility to AI. Their dispute revolves around the fact that AI will never have autonomous will, because it can never do something for which it has not been programmed [12]. AI has a defined set of actions that it cannot act outside of. One criticism of this view is that in this sense, humans are not moral agents either, because their beliefs, goals, and desires are not completely autonomous and are the product of a set of factors consisting of culture, technology, environment, education, and brain chemistry.

- ***Group Four: Artificial Intelligence Has Agency but No Responsibility***

Some researchers believe that although artificial intelligence has agency, it cannot accept responsibility for its actions. Because it does not have freedom and consciousness - which are the prerequisites for responsibility from its point of view - artificial intelligence is influential in shaping human actions and is not simply a neutral tool, and in this sense, it is neither responsible nor irresponsible; rather, it is involved in human responsibility. According to this group, humans are also responsible for its errors because they have entrusted decision-making and the power to do the work to artificial intelligence. In order to prevent such situations as much as possible, humans are obliged to develop artificial intelligence in the most responsible way possible and to manage and predict its risks in a priori.

Other people who fall into this category believe that in order to resolve moral paradoxes, we must accept a mindless morality that avoids issues such as free will and intentionality. There is still controversy in the philosophy of mind about such concepts, and these same debates have entered the field of artificial intelligence and have inappropriately affected this field. Moreover, conventional philosophy and ethics, by attributing moral agency exclusively to humans, place a great deal of responsibility on human agents [13]. Of the four perspectives discussed, the last approach, that is, acknowledging agency for artificial intelligence and not assigning responsibility to it, seems to be the most acceptable. Because artificial intelligence has agency in the causal sense of the word and is active. However, in our opinion, the purpose of assigning moral responsibility to an entity is to be able to punish it when it makes a mistake. Since artificial intelligence does not perceive pain or pleasure and does not have mental states of will and consciousness (at least for now), it cannot take responsibility for its own mistakes.

Strategies for Responsible Development of AI Technologies

The responsibility gap places humans in a dilemma: either they should accept the development of AI technologies despite the responsibility gap, or they should stop their development because of this gap. Given that attributing responsibility to AI cannot help its responsible development, various philosophers have tried to offer strategies for the development of AI technologies that create a way out of this dilemma. The basis of these strategies is the assumption that at present, and while humans can still have agency and responsibility for the path of technological development, they should use all possibilities to minimize the risks of such technologies. These strategies impose requirements for the path of AI development that reduce the responsibility gap as much as possible, and help make AI more predictable and controllable. Some of these strategies are discussed below:

A. *Meaningful Human Control*

Researchers distinguish four different interpretations of the accountability gap: punitiveness, moral accountability, public accountability, and proactive accountability [14].

Punitiveness: The punitive aspect of responsibility comes into play when the use of technology causes an undesirable event. In this case, the public and the victims of the incident need to know who was at fault for the incident. Punishing perpetrators of the incident, in addition to compensating victims, also has a deterrent effect and is a means of establishing and reinforcing shared social norms. Punitiveness is not a phenomenon that emerged with the formation of artificial intelligence, but originally goes back to the problem of “many hands.” But artificial intelligence technologies—and especially their learning type—have made this gap more tangible. In cases where multiple actors are involved in an incident, it is possible that none of the actors will take responsibility or that some actors will delegate it to others in an unauthorized manner [14].

Moral accountability: Moral accountability refers to the actors’ awareness of the moral dimensions of their actions. This aspect of responsibility requires the individual to reflect on the relationship of the action to moral values and to be able to explain this relationship to others. By clarifying the relationship of the action to the individual’s values and beliefs, the individual gives meaning to his actions, becomes more aware of his agency and role, and accepts responsibility for his behaviors [15]. In the case of AI technologies based on deep learning, the lack of transparency in decision-making processes prevents the individual using the program from being able to properly analyze it and meaningfully understand his role in this process and his responsibility for the consequences of what happened.

Public accountability: Another dimension of accountability is public accountability. Public accountability is usually about management and policy decisions, and since AI can appear in a managerial position and enforce the law, the issue of public accountability makes sense for it. In the accountability process, a dialogue is formed between the perpetrator of an event and representatives of the public, in which the accountable person is challenged and his decisions and responses are judged. However, in the case of AI technologies, a public accountability gap is seen in several ways. One aspect of the problem goes back to the phenomenon called the "technical black box." This means that the causal processes and algorithms of these systems are sometimes so complex that even experts do not have sufficient mastery over them and, consequently, cannot adequately explain these processes to the public. But another aspect of the gap goes back to the existence of the "institutional or legal black box." When AI technologies are deployed at the management level, decision-making processes involve institutions and individuals who, for various reasons, are not subject to public accountability. For example, the development of information technologies is usually entrusted to private companies and tech giants such as Google, which are reluctant to disclose errors in their systems. In addition, data is exchanged in large volumes and between multiple institutions, making it difficult to determine who should be held accountable in the event of a problem [14].

Active accountability: Active accountability is the fourth aspect of accountability threatened by AI. Unlike passive accountability, which looks back at events, active accountability involves measures taken by individuals and institutions to better fulfill their roles and act in accordance with defined norms. In some perspectives, this idea is presented by distinguishing between "backward" and "forward" approaches. For example, when we examine who is at fault in a self-driving car accident, we are looking backwards. When we look forward, we are looking at how the technology and its social environment evolved to figure out how to prevent the accident. So, in developing AI responsibly, we can in a sense seek to reflect on the unintended consequences and ethical implications of AI before these technologies are widely used.

Considering the different dimensions of accountability, an approach called "meaningful human control" has been proposed to solve the accountability gap. In this approach, firstly, the focus is on the responsibility and role of human agents, and secondly, instead of simply adding legal appendices, the way of institutional expansion and development of AI technologies is considered. This approach introduces two necessary conditions for the development of AI technologies: Tracking for institutional-technical transparency and creating traceability.

Condition One: Tracking

The tracking condition is about the nature of control relationships and the creation of features in the socio-technical systems of AI, which are necessary to maintain human responsibility in such a system. The feasibility of human control does not only require the creation of new technical features to increase technical transparency, but also requires social, institutional and legal arrangements that guarantee institutional and political transparency about control roles. In fact, according to the tracking condition, the path of technology development should be such that the role of different institutions in the development process and its relationship to the intended values are transparent. This systematic specialization allows for the accountability of the system. Meeting this condition requires, first, a map of all the actors involved in the design, control, regulation, and use of a system, and then an analysis of the relationship of these actors to the intended reasons, values, and purposes. This condition actually facilitates the explainability of AI technologies, which Coeckelbargh also emphasizes as an essential component of responsibility [15]. It also enhances active accountability, since the transparency of processes and roles clarifies the values and norms of each part and enables commitment to them.

Condition Two: Tracing

The tracing condition is about the possibility of tracing the system outputs and reaching a human agent who can be held accountable for the system's performance (or part of its performance). A person who has played a role in the design, development, and usage chain, and at the same time has sufficient information about the system's capacities and limitations, also has the necessary moral awareness and has the capacity to be legally accountable for the system's behavior. Although traditionally the agent responsible for an action is considered to be someone who has a causal role in causing that action; however, according to the tracing condition, in the case of AI systems we cannot directly and exclusively identify causal agency. Because in complex socio-technical systems, there are multiple and interacting causal chains, it is not possible to find an agent who has both a technical role and sufficient moral awareness of the system's outputs. According to the tracing condition, such a position should be defined in socio-technical systems related to AI; A position in which an individual has sufficient technical and ethical knowledge of the system processes, and the capacity to change and direct them. Therefore, he or she can reasonably be held accountable. The goal is to define this position to have a fair distribution of moral culpability that prevents two types of errors: that someone is punished without being able to avoid the error, and that someone is not punished for avoidable errors. Defining such a position can reduce the gap between moral accountability, public accountability, and culpability [14].

B. Triple agency model

One of the models that can partially fulfill the two conditions of tracking and routing is a model that distinguishes three types of agency from each other: 1) Causal agency, in which, when entities enter into causal relationships, they act on each other, interact with each other, and create changes on each other, and as a result, they have a causal effect. 2) Voluntary agency, in which the agent has a will and his will leads to an action, and this action is related to responsibility through a complex relationship. 3) Triple agency, which is an agency beyond the sum of individual actions; for example, when a human achieves a goal in cooperation with artificial intelligence.

In this model, three types of agents are distinguished: 1) The user, who wants to achieve a goal and entrusts this task to the designer. 2) The designer, who designs the artificial to achieve the goal. 3) The artifact, which produces the causal effect necessary to achieve the goal. In order to determine how responsibility for various consequences should be distributed among willing agents, the behaviors of human components can be examined in the form of these triplets. According to Johnson and Verdicchio, assigning responsibility to artificial intelligence is neither meaningful nor beneficial. Because artificial intelligence is not capable of performing voluntary and intentional actions and does not have voluntary agency nor responsibility. As a result, it is humans who, by entrusting complex tasks to artificial intelligence, must also take responsibility for unintended consequences. However, to determine the degree of responsibility of each individual, we need to consider the role of technological components.

As artificial intelligence advances, the gap between human decision-making and the performance of artificial intelligence increases, but according to Johnson and Verdicchio, the triplets model helps us to trace agency and responsibility in different processes [16].

C. Relevance and Collective Accountability

While the conditions of tracking and tracing are useful in bridging the accountability gap, there is another dimension of accountability that goes beyond these two. This dimension is called relevance of accountability. When we think about the relationship between technology developers and their users, we should not consider users as merely passive elements. Granting a role and agency to the user community is important for several reasons. First, given the importance of public and ethical accountability, providing explanations occurs in an interactive process between technology developers and their potential users. This means that users have a role in determining the instances of explanation and the quality of such explanations. Therefore, for accountability and explainability to be achieved, it is necessary for the general public to have a good understanding of the issues related to AI technologies, and their potential impacts on their lives and interests. In the absence of public awareness and a sense of need, an important part of the motivation for making development processes transparent will be lost.

But the importance of engagement with audiences is not limited to the issue of explainability. In the functioning of AI technologies that are capable of learning, part of the development of the technology actually occurs through interaction with the environment (i.e., users). As those who provide input to AI systems, users play a role in how these systems are formed and behave.

Also, the development of technology and technological practice is a collective act that entails collective responsibility. An act in which not only a set of formal elements related to the development of the technology play a role; but also, the general public and the intellectual and cultural characteristics of the society play a role. An AI technology may have a value bias that reflects the institutionalized bias in the whole society and realizes values that are ingrained in the culture of the society. We can therefore add a third condition to the list of essentials for the responsible development of AI technologies: public engagement.

According to this condition, along with the technical and institutional development of technology, it is necessary for the user community to be involved in the technology development process. In this interactive process between developers and users, the necessary information about the developing system and its potential possibilities is given to the users. In this way, firstly, the users' minds become sensitive to the subject and new questions are formed in it, which enriches the process of explanation. They also become familiar with the role of various institutions in technology development, which creates the possibility of demanding and questioning in them. Secondly, they become aware of the role that they themselves can have in this process and the responsibilities that follow from this role. In this case, it arouses a sense of active responsibility in citizens.

Different methods and tools have been proposed for involving citizens in technology development:

- Formation of focus groups in which interested individuals discuss their opinions and preferences;
- An interactive process of scenario analysis in which a group identifies key issues, designs and examines different scenarios, and tests them using the opinions of different individuals;
- Forming a group of diverse participants and pooling their knowledge to develop and explore ideas and policies;
- Forming panels of informed citizens who are representative of public beliefs [17].

CONCLUSION

Artificial intelligence technologies are advancing humanity towards an unknown and ambiguous future at an ever-increasing pace. The complexity of these technologies and their learning nature make their development path somewhat out of control and unpredictable, and considering the irreversibility of the effects of these technologies, the necessity of their responsible development seems obvious.

Various answers have been given to the issue of the responsibility gap. Some believe that society can attribute responsibility for the outputs of these technologies to them. But others oppose this idea and believe that while artificial intelligence technologies have agency, this does not mean they have responsibility, and attributing responsibility to them does not help control their outputs by human agents. Accordingly, only the human actors involved in the development of these technologies can be held responsible, and they should be expected to use their freedom and awareness to shape the path of technological development in a way that prevents unpleasant and unethical incidents. In general, the three principles of routing, tracking, and engaging public opinion and attention to public emotions in policymaking can be useful as three effective strategies in the responsible development of artificial intelligence technologies.

ETHICAL CONSIDERATIONS

Ethical issues (such as plagiarism, conscious satisfaction, misleading, making and or forging data, publishing or sending to two places, redundancy and etc.) have been fully considered by the writers.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests.

REFERENCES

1. Rudy-Hiller F. The epistemic condition for moral responsibility. 2nd ed. USA/Stanford: Metaphysics Research Lab, Stanford University. 2022.
2. J. M. &. R. M. Fischer, Responsibility and control: A theory of moral responsibility, Cambridge: Cambridge University Press, 1998
3. Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 2004; 6(3): 175-183. Doi: 10.1007/s10676-004-3422-1
4. Himma K E. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics and Information Technology, 2009; 11(1): 19-29. Doi: 10.1007/s10676-008-9167-5
5. Haselager W F. Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition, 2005; 13(3): 515 – 532. Doi: 10.1075/pc.13.3.07has
6. Rodogno R. Social robots, fiction, and sentimentality. Ethics and Information Technology, 2016; 18(4): 257–268. Doi: 10.1007/s10676-015-9371-z
7. List C, Pettit P. Group agency: the possibility, design, and status of corporate agents. 2nd ed. London: Oxford University Press. 2011.
8. Dennet D. Brainchildren: Essays on designing minds (representation and mind). USA: MIT Press, 1998.
9. Sullins J P. When is a robot a moral agent? International Review of Information Ethics, 2006; 6(12): 23-30.
10. Latour B. Pandora's hope. Essays on the reality of science studies. 1st ed. USA: Harvard University Press, 1999.
11. Dennett D. When HAL kills, who's to blame? in Stork, David, HAL's Legacy: 2001's Computer as Dream and Reality, MIT Press, 1998. Doi: https://doi.org/10.7551/mitpress/3404.003.0018
12. Bringsjord S. Robots: The future can heed us. AI and Society 2007; 22(4):539-550. Doi: 10.1007/s00146-007-0090-9
13. Floridi L, Sanders J W. On the morality of artificial agents. Minds and Machines,2004; 14(3): 349-379. Doi: 10.1023/B:MIND.0000035461.63578.9d
14. Desio F, Mecacci G. Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philosophy and Technology, 2021; 34(4):1-28. Doi: 10.1007/s13347-021-00450-x
15. Coeckelbargh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and Engineering Ethics, 2020; 26(2): 2051-2068. Doi: 10.1007/s11948-019-00146-8
16. Johnson D, Verdicchio M. AI, agency and responsibility: the VW fraud case and beyond. AI & Soc, 2019; 34: 639–647. Doi: 10.1007/s00146-017-0781-9
17. Roeser S, Pesch U. An emotional deliberate approach to risk. Science, Technology and Human Values,2016; 41(2): 274-297. Doi: 10.1177/0162243915596231