https://doi.org/10.25038/am.v0i28.616

Bojan Blagojević
Faculty of Philosophy, University of Niš, Serbia

Raising Skynet: Moral Status of AI and Perspectives of Teaching Ethics to AI

Abstract: Ever since MIT's *Norman* and *Deep Empathy*, the profound impact of unexamined biases in AI research has become apparent. Beyond extreme cases like the 'psychopathic' *Norman*, broader concerns surface, ranging from accusations of AI being labeled 'too liberal' or 'leftist' to claims of it being 'racist' or 'sexist'. This paper seeks to move beyond conventional narratives and focuses on the overlooked responsibility stemming from the moral status of AI. Central to this exploration is the examination of three key aspects.

Personhood: the concept of AI personhood involves the possibility of AI entities as individuals with distinct rights and moral considerations. Determining the criteria for personhood in AI and how it aligns with or diverges from human personhood establishes the foundation for ethical frameworks.

Embodiment: in the realm of AI, embodiment raises questions about the nature of AI's interaction with the physical world. The extent to which AI is grounded in physical form or exists in a virtual domain impacts its ethical considerations and is crucial for establishing its moral status.

Sensitivity to pain/pleasure: the capacity of AI to perceive and respond to pain and pleasure, introduces complex ethical implications. Assessing them requires an exploration of the responsibilities tied to AI's potential to influence or be influenced by positive and negative experiences.

These considerations contribute to a nuanced understanding of the moral status of AI. By addressing the intricacies of these aspects, we aim to provide an outline for teaching ethics to AI, having in mind that no existing artificial system meets even the minimum criteria for moral agency or moral patience.

Keywords: AI ethics; moral status; embodiment; personhood; education.

Introduction

As AI continues to permeate both private and public spheres, expanding its influence across diverse knowledge domains and human practices, it gives rise to a growing array of speculations, philosophical debates, and political dilemmas. This paper will focus on a specific subset of ethical issues related to AI: the unexamined biases in AI research and the diverse range of human responsibilities associated with it. Bias is often viewed as having undesirable consequences for individuals, groups, or even entire

Author's contact information: bojan.blagojevic@filfak.ni.ac.rs

societies. However, this paper aims to explore the possibility that such biases may also have detrimental effects on AI itself. To achieve this, I will begin by examining the types of biases present in AI research. Next, I will address the question of AI's moral status, exploring both its potential as a moral agent and a moral patient, along with the key aspects tied to these concepts. Finally, I will present and assess various perspectives concerning the development of so-called *moral machines,* analyzing how this endeavor could relate to the process of moral education or upbringing and outlining the characteristics of an educational ethics suitable for AI moral development.

To mirror the world or to improve it?

One of the probably most famous (or most notorious) examples of bias concerning AI systems is the use of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) by courts in the US "to assess the likelihood of a defendant becoming a recidivist."[1] It was revealed that the COMPAS algorithm could predict a convicted criminal's likelihood of reoffending. However, an independent investigation uncovered that when the algorithm's predictions were incorrect, its errors disproportionately impacted black offenders. These individuals were more often denied parole due to biased data embedded in the system.[2]

Bias can emerge at various stages of design, testing, and application. Focusing on design, it may stem from the selection of the training dataset, the dataset itself (if it is unrepresentative or incomplete), the algorithm, the data provided to the algorithm after training, or decisions influenced by spurious correlations. Bias can also arise from the composition of the team developing the algorithm or from broader societal influences. Additionally, research has shown that machine learning systems can absorb bias from textual data sourced from the internet, as such data inherently reflects human culture, including its biases. Let us take *Norman* as another example:

> Norman, world's first psychopath AI. Norman was inspired by the fact that the data used to teach a machine learning algorithm can significantly influence its behavior. So when people say that AI algorithms can be biased and unfair, the culprit is often not the algorithm itself, but the biased data that was fed to it. The same method can see very different things in an image, even 'sick' things, if trained on the wrong (or, the right!) data set. Norman suffered from extended exposure to the darkest corners of Reddit, and represents a case study on the dangers of artificial intelligence gone wrong when biased data is used in machine learning algorithms.[3]

[1] "COMPAS (software)", Wikipedia, The Free Encyclopedia, accessed January 28, 2025, https://en.wikipedia.org/wiki/COMPAS_(software).

[2] Julia Angwin et al., "Machine Bias", ProPublica, accessed January 28, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] "Project Norman", MIT Media Lab, accessed January 28, 2025, https://www.media.mit.edu/projects/norman/overview/.

Now we are aware that there are two distinct, albeit connected issues at hand. The first concerns the connection between bias inherent in AI systems and its societal influence. Addressing bias in AI is not solely a technical challenge; it is fundamentally a political and philosophical one. It raises questions about the kind of society and world we envision, whether it should be transformed, and, if so, what methods of change are both just and fair. On one hand, AI seems to perpetuate and even amplify the biases embedded in the data it is trained on, further deepening the disadvantages experienced by historically marginalized groups. On the other hand, debates continue over the most appropriate and just methods for addressing existing prejudices and discrimination. For instance, there is an ongoing debate about whether positive discrimination or affirmative action, aimed at counteracting bias by favoring disadvantaged individuals or groups, is a just approach. Thus, we are dealing with two different conceptions of the role of AI. One of them argues that AI should merely mirror the world as it is, while the other envisions it as a tool to be used in order to reshape our perception and perhaps improve the world. Even if we agree on the latter conception, we are facing other crucial difficulties: philosophers and society at large have yet to reach a consensus on what constitutes perfect justice or fairness. Does true justice necessarily require complete elimination of all bias? It is worth questioning whether a perfectly unbiased algorithm is even achievable, and, if it is, whether it would be desirable. As Mark Coeckelbergh notices, "It is not clear if bias is avoidable at all or even if it should be avoided, and, if so, at what cost it should be avoided. For example, if changing the machine learning algorithm in order to decrease the risk of bias makes its predictions less accurate, should we change it? There may be a trade-off between effectiveness of the algorithm and the countering of bias."[4]

There is another option that we should consider. Perhaps these difficulties arise from *human* decision making. Perhaps AI could offer insights into human nature and society by exposing our biases, while discussions around AI ethics might shed light on existing social and institutional power imbalances. Coeckelbergh phrases this as the idea that there is no "AI in itself".[5] The idea that "AI in itself" does not exist stems from the fact that the technology is inherently social and human. AI is not just about the technology itself but also about how humans use, perceive, experience, and integrate it into broader socio-technical systems. This perspective is crucial for ethics, centered on human decision-making, and highlights the importance of incorporating historical and socio-cultural contexts. If we expect AI systems such as COMPAS to produce outcomes that we could consider just, it would likely require them to 'understand' such contexts and the power-play contained within them. Additionally, they would need to be trained that there are certain questions that may always be controversial and open for debate.

The second issue mentioned above concerns the fate of *Norman*. Trained solely on one type of data – posts from a Reddit community featuring graphic videos of

[4] Mark Coeckelbergh, *AI Ethics* (MIT Press, 2020), 131.

[5] Coeckelbergh, *AI Ethics*, 80.

people dying – *Norman* interpreted Rorschach inkblots entirely as depictions of death and execution. Not surprisingly, no one batted an eye when they pulled the plug on old *Norman*. After all, 'he' was merely an algorithm. However, I want to delve deeper into this phenomenon – not into *Norman* specifically, but into our broader reaction to the entire situation. If we assert that Norman wasn't wronged or mistreated, as most of us are no doubt inclined to do, it implies that we do not attribute any moral status to 'him'. It is worth considering what criteria would need to be met for us to ascribe moral status to an AI.

Moral status of AI

Most ethical questions surrounding AI rarely address AI itself directly. Instead, they center on humanity's future, often imagining risks arising from an unprecedented context. In grim dystopian scenarios, AI is portrayed either as a mighty tool wielded by the power-hungry or as an autonomous force intent on eradicating 'inferior' humanity. However, beyond these speculative narratives, there are many realistic and pressing issues that the ethics of AI seeks to address. I will focus on the questions concerning the moral status of AI.

The term "moral status" encompasses two types of questions. The first involves what an AI is capable of doing from a moral perspective – essentially, whether it can possess moral agency, and if so, whether it can qualify as a full moral agent. What does this entail? It seems clear that the actions of AI systems today already have moral consequences. Most would agree that AI currently exhibits a 'weak' form of moral agency, comparable to the role of modern cars. However, as AI becomes increasingly intelligent and autonomous, the question arises: can AI develop a stronger form of moral agency? Should it be granted – or could it acquire – the capacity for moral reasoning, judgment, and decision-making? For instance, should self-driving cars that rely on AI be considered moral agents, and if so, to what extent?

In addition, questions about 'moral status' can also pertain to how we ought to treat AI. If a highly intelligent artificial entity were to be developed in the future, would we be obligated to grant it rights, even though it is not human? Philosophers refer to this as the issue of moral patience. This question shifts the focus from the ethics within or by AI to our ethical responsibilities toward AI. In this context, the AI becomes a subject of ethical consideration, rather than being viewed as a potential ethical agent.

We can compare this question to the question of the moral status of animals. Today, many people believe that animals hold moral significance, but this wasn't always the case. It seems clear that we were mistaken about animals in the past. If many people today view AIs as nothing more than machines, could they be making a similar error? For instance, would superintelligent AIs warrant moral status? Should they be granted rights, or is it potentially dangerous to even entertain the idea that machines might possess moral status?

When it comes to the status of moral agents, there are opinions that machine morality is possible and quite desirable. "*(M)achine ethics* is concerned with giving *machines* ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making."[6] This 'giving' of principles/procedures should be a result of interdisciplinary cooperation. However, from an ethicist's point of view, this very cooperation seems much less likely when we encounter the following statement: "Ethicists must accept the fact that there can be no vagueness in the programming of a machine, so they must sharpen their knowledge of ethics to a degree that they may not be used to."[7] Although widely utilized by various ethical frameworks throughout the history of philosophy, this 'top-down' approach to constructing morality through strict principles and procedures is, at best, highly questionable. Coeckelbergh notices an additional problem concerning the complexity of the moral phenomenon, specifically the role of emotions in morality:

> The entire project of building 'moral machines' by giving them rules is based on mistaken assumptions regarding the nature of morality. Morality cannot be reduced to following rules and is not entirely a matter of human emotions – but the latter may well be indispensable for moral judgment. If general AI is possible at all, then we don't want a kind of 'psychopath AI' that is perfectly rational but insensitive to human concerns because it lacks emotions.[8]

For these reasons, we might reject the concept of full moral agency for AI entirely or adopt a middle-ground approach: providing AIs with a form of morality, but not full moral autonomy. Wallach and Allen refer to this as "functional morality"[9], where AI systems are equipped with some ability to assess the ethical consequences of their actions. This approach is particularly relevant for self-driving cars, as these vehicles are likely to encounter situations requiring moral decisions when there is no time for human input or intervention. Some of these scenarios involve moral dilemmas, such as the classic trolley problem. What should the car decide? It seems we will need to make these moral choices in advance and ensure that developers program them into the cars. Alternatively, we could design AI cars capable of learning from human decision-making. However, this raises questions about whether providing AIs with fixed rules is an accurate way to represent human morality – assuming morality can even be 'represented' or reproduced – and whether trolley dilemmas truly capture the essence of moral experience. From a different perspective, one might also question whether humans themselves are effective at making moral decisions. If that is the case, why should AI imitate human morality at all?

[6] Michael Anderson and Susan Leigh Anderson, eds., *Machine Ethics* (Cambridge University Press, 2011), 1.

[7] Anderson and Anderson, *Machine Ethics*, 3.

[8] Coeckelbergh, *AI Ethics*, 52.

[9] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2009), 39.

Another area of controversy involves the moral patience of AI. If we were to create a superintelligent AI, would it be morally acceptable to switch it off or 'kill' it? Taking a more immediate example: is it ethical to kick a robotic AI dog? There is already research that shows that people today empathize with robots and hesitate to 'harm' or 'kill' them, even if these robots lack AI capabilities.[10] Humans appear to need very little from artificial agents to project qualities like personhood or humanness onto them and to feel empathy for them. As these agents become more advanced, incorporating AI and appearing even more human- or animal-like, the question of their moral patience becomes increasingly pressing. How should we respond to those who empathize with an AI? Are they mistaken in doing so? A common and intuitive stance is to assert that AIs are simply machines and that people who empathize with them are misguided in their judgments, emotions, and moral experiences. At first glance, it seems we owe nothing to machines – they are objects, not people. Those who hold this view might agree that if AIs ever became conscious or capable of having mental states, they would deserve moral status. However, they argue that such conditions are not met today. The problem with this position, however, is that it fails to explain or justify the moral intuitions and experiences many people have when they feel that 'mistreating' an AI is wrong, even if the AI lacks consciousness or sentience.

One possible justification for this intuition comes from Kant, who argued that it is wrong to harm a dog – not because it violates a duty to the dog itself, but because it "damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind."[11] A virtue ethics perspective offers a similar, indirect argument: mistreating an AI is wrong, not because the AI itself is harmed, but because such behavior harms our moral character. It fails to make us better people.

But if we were to consider moral relevance of AI in its own right, where would that take us? Let's think about how we can determine whether an AI truly possesses morally relevant properties. Are we even certain about these properties when it comes to humans? A skeptic might argue that we are not. Yet, despite this lack of epistemological certainty, we still attribute moral status to humans, often based on appearance alone. This same tendency could likely extend to AIs in the future, especially if they were to exhibit human-like appearances and behaviors (as android AIs). Looking more closely at how humans actually assign moral status reveals that factors like social relationships play a significant role. For instance, we don't treat our dog kindly because of deep moral reasoning about the dog's nature; we do so because we already share a social bond with it – it is our pet and companion. Similarly, when we give our dog a personal name, we confer a moral status upon it that we wouldn't attribute to the nameless animals we consume. In a similar vein, we could argue that the moral

[10] Yutaka Suzuki, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki, "Measuring Empathy for Human and Robot Handpain Using Electroencephalography," *Scientific Reports 5*, (November 2015) 1–9, https://doi.org/10.1038/srep15924. Kate Darling, Nandy Palash, and Cynthia Breazeal, "Empathic Concern and the Effect of Stories in Human-Robot Interaction," *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication*, (New York, IEEE, 2015), 770–775, https://doi.org/10.1109/roman.2015.7333675.

[11] Immanuel Kant, *Lectures on Ethics* (Cambridge University Press, 1997), 212.

status of AIs will be determined by human beings and shaped by how these entities are integrated into our social lives, our language, and our broader cultural practices.

Thus, while we typically require certain prerequisites to ascribe the status of a moral agent – such as responsibility, consciousness, free will, and the ability to form intentions – AI might still be capable of moral patience without meeting any of these criteria. It would be useful, then, to look into these criteria.

What does moral status require?

Many ethical questions surrounding AI inherently involve comparisons – both implicit and explicit – between humans and AI. In fact, AI's success or failure is often evaluated based on how well it replicates or surpasses human abilities. However, such comparisons are inevitably shaped by value judgments; for instance, asserting that AI has outperformed humans in an intelligence-based task depends on underlying assumptions about what qualifies as 'better'. Moreover, these comparisons are always influenced by broader conceptual frameworks and preconceptions about both human nature and the role of machines. While perspectives such as transhumanism and post-humanism may offer valuable insights into the matter, we will stick with a humanistic framework for the time being.

What are the most striking distinctive characteristics of the human condition and experience when it comes to action, and subsequently, morality? The uniqueness of human individuality, along with the narrative structure of our lives – marked by strength, weakness, resilience, despair, and the inherent limitations of our physical existence – are some of the essential considerations. Can such qualities ever be ascribed to machines? Furthermore, key principles such as freedom and autonomy, transparency and accountability, responsibility, justice and fairness, beneficence and non-maleficence, privacy, trust, sustainability, dignity, and solidarity must also be considered. However, before we rush to conclude that these qualities and principles cannot be attributed to AI, we should first acknowledge that not all human beings inherently possess them either. The concept of a 'person' typically refers to a being capable of both moral agency and moral patience – that is, a being to whom moral status can be attributed. Still, as Paula Boddington warns, we must be aware that the concept of person is

> an incredibly complex issue, since there are many ways of understanding the concept of a person, let alone the ethical significance of the category of the person. […] It is challenging to write an account that goes: 'This is what a person is, now let's look at the ethical implications for AI', because determining the category of a person and spelling out precise criteria for personhood is not only controversial, it is so intimately tied up with value issues.[12]

[12] Paula Boddington, *AI Ethics* (Springer, 2023), 320.

Bearing this in mind, I will limit myself to a moral notion of the person, that is often used to denote a being of particular standing and worth. In some uses, 'person' serves as a term to signify that a being holds certain moral claims, that is to say, one to which we can ascribe moral patience. Other uses set persons apart from other agents – who may act but do not qualify as persons – by precisely their capacity for moral reflection, self-awareness, and the ability to evaluate their own beliefs and actions, that is to say, ascribe them moral agency. I will use a specific moral conception of person and personhood provided by Daniel Dennett and examine its ethical implications in the context of AI.[13]

Dennett examines six concepts that claim to be necessary conditions for personhood and examines (among other things) whether they jointly constitute a sufficient condition for personhood. The concepts are: rationality, consciousness of mental or intentional content, a certain treatment or status ascribed by others, ability to reciprocate that treatment, verbal communication, and self-consciousness (i.e. consciousness of itself *as* a person).[14] Persons are not *a priori* limited to human beings, or any species, for that matter. The idea that a person must be capable of reciprocating in their treatment of others suggests a form of embodiment, as it requires the ability to interact and take action in the world. Similarly, the requirement for verbal communication implies a capacity for physical or functional engagement.

It seems that the requirement of self-consciousness rules out AI as a candidate for a person. The ability to be aware of one's own thoughts and emotions – a higher-order form of consciousness – may be a crucial component of personhood as it enables the reflection on one's plans, actions, and internal states necessary for making judgments, decisions, and pursuing goals. However, does this necessarily require a distinct, unique sense of self? And how might such a "self" be conceptualized? Perspectives on the self are likely shaped by cultural and ideological frameworks, influencing how it is understood and its relationship to others.

Similarly, there is a *Catch-22* in the notion that personhood depends on being recognized as a person by others. If humans refuse to grant AI this recognition, it could meet all other criteria for personhood and still be denied that status.

The question of embodiment is now crucial for us, particularly as it relates to the individuation of persons – an essential issue in ethics. Our human bodies naturally define clear boundaries around our identities as individuals, making embodiment deeply significant to our sense of personhood. A key aspect of this is the human face, which is remarkably expressive and fundamental to social interaction. Faces serve a dual function: they connect us as members of a shared humanity, enabling communication as a social species, while also distinguishing our unique individuality. How does this work with AI? Even if we were to attribute moral status to an AI that exists diffusely – spread across different parts of the Internet and engaging with us through language and communication – this would still constitute a form of embodiment.

[13] Daniel Dennett, *Brainstorms* (MIT, 2017), 278–306.

[14] Dennet, *Brainstorms*, 289–91.

Moreover, sustained existence over time appears essential for ascribing goals, preferences, or purposes, as these require continuity and persistence to be meaningful. However, the role of the face would require that AI becomes capable of a sense of itself as a unique person and ascribe value to that fact.

Value judgements serve as a basis of another important criterion for attaining moral status: sentience. What is the role of sentience in moral consideration? Does a robot need to have experiences (especially of pleasure and pain) in order to be a proper recipient of moral concern? Even if we were to attribute sentience to a machine, sentience alone might not be sufficient to determine how we should treat it. Moral judgments attached to sentient experiences, as well as the goals and values associated with them, appear to be essential. In discussions of personhood, the right to life is often tied to the desire to continue living. Similarly, the wrongness of inflicting suffering on sentient beings is typically based on the quality of their experience – specifically, the negative nature of pain. This suggests the presence of an experiencing subject with at least the rudiments of a value system, a motivation to avoid the negative and seek the positive. An alternative perspective might be that, just as our moral responsibilities toward animals differ from those toward humans – and even vary across species – we may need to develop a new ethical framework specifically suited to machines.

There are numerous other issues that fall beyond the scope of this discussion. The issues addressed so far establish a framework for tackling one final question: can there be an AI-educational ethics and what it would look like.

"Bringing up Baby-bot"

It's time to say it outright: the jig is up. No existing artificial system meets even the minimum criteria for moral agency or moral patience. The AI systems developed to date lack anything that could be considered moral insight, compassion, or even basic common sense and decency, despite their potential usefulness in achieving morally desirable outcomes. Most of current research in AI ethics blends practical considerations about what we can realistically expect from existing or near-future AI with more speculative discussions about what an ideal AI might one day become. This includes the possibility of AI developing sentience, compassion, emotions, and other fundamental aspects of human nature that shape our moral responses. Still, as Coeckelbergh notes, one of our tasks is to "investigate AI policy for the near future."[15] If the near future unfolds as envisioned by Wallach and Allen, it is worth considering: if we can "train" machines to be moral, is there a right or a wrong way to do so that extends beyond its impact on humanity alone?

Moral Machines by Wendell Wallach and Colin Allen provides an in-depth exploration of why we might need machines to be moral. The book outlines two possible approaches: a top-down method, which involves defining a moral framework and embedding it into machines, and a bottom-up method, which seeks to integrate

[15] Coeckelbergh, *AI Ethics*, 29.

ethical considerations into machines through a developmental process, mirroring how humans acquire moral judgment and behavior.

> (A) top-down approach takes an ethical theory, say, utilitarianism, analyzes the informational and procedural requirements necessary to implement this theory in a computer system, and applies that analysis to the design of subsystems and the way they relate to each other in order to implement the theory.
> In bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and learns and is rewarded for behavior that is morally praiseworthy. There are various models for bottom-up acquisition of moral capabilities. Childhood development provides one model. Evolution provides another bottom-up model for the adaptation, mutation, and selection of those agents best able to meet some criteria for fitness. Unlike top-down ethical theories, which define what is and is not moral, in bottom-up approaches any ethical principles must be discovered or constructed.[16]

By referencing the childhood development model, Wallach and Allen touch on a crucial challenge for near-future AI. When examining beings that lack moral agency but are granted moral patience, we see significant distinctions. Animals, for instance, are considered moral patients and are treated accordingly, with the understanding that they will never develop moral agency. In contrast, children are nurtured and educated with the expectation that they will eventually attain moral agency. This raises a potentially significant question: if AI is to be treated as a moral patient, should it be viewed more like an animal – permanently lacking agency – or more like a child, with the potential to develop into a moral agent over time? Another option is to refrain from treating AI as a moral patient altogether, given that it does not meet the necessary criteria. Some might argue, however, that the only way for AI to eventually fulfill these criteria is to treat it as if it already does. Do we have reason to believe this is the case? If so, we have basis for an AI-educational ethics. A different question is whether we actually want an Artificial Moral Agent, as Wallach and Allen call it, since its creation would bring many of the same challenges we face in raising and educating new generations, with the horrific additional risk of *the unknown*.

Generating out-put

What I have presented is only a limited discussion. I left aside alternative perspectives such as transhumanism and posthumanism, which often argue that human ethics has reached its limits and ought to be replaced with something more advanced, or simply different. It may be futile to debate which dystopian scenario is more

[16] Wallach and Allen, *Moral Machines*, 80.

terrifying: a superintelligent AI treating humans as humans have treated animals they deemed inferior, or a singularity in which AI operates in ways entirely beyond human comprehension. Engaging in a step-by-step process of teaching morality to AI may help us to distance ourselves from apocalyptic speculations and reflect on our nature and our values.

References

Anderson, Michael, and Susan Leigh Anderson, eds. *Machine Ethics.* Cambridge University Press, 2011.

Angwin, Julia et al. "Machine Bias". ProPublica. Accessed January 28, 2025. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Boddington, Paula. *AI Ethics.* Springer, 2023.

Coeckelbergh, Mark. *AI Ethics.* MIT Press, 2020.

Darling, Kate, Nandy Palash, and Cynthia Breazeal. "Empathic Concern and the Effect of Stories in Human-Robot Interaction." *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication*, (2015), 770–775. https://doi.org/10.1109/roman.2015.7333675.

Dennett, Daniel. *Brainstorms.* MIT, 2017.

Kant, Immanuel. *Lectures on Ethics*. Cambridge University Press, 1997.

MIT Media Lab. "Project Norman". Accessed January 28, 2025. https://www.media.mit.edu/projects/norman/overview/.

Suzuki, Yutaka, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki. „Measuring Empathy for Human and Robot Handpain Using Electroencephalography", *Scientific Reports* no. 5, (2015): 1–9. https://doi.org/10.1038/srep15924.

Wallach, Wendell, and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, 2009.

Wikipedia, The Free Encyclopedia. "COMPAS Software". Accessed January 28, 2025. https://en.wikipedia.org/wiki/COMPAS_(software).

Article received: December 12, 2024
Article accepted: February 1, 2025
Original scholarly article