AN ETHICAL APPROACH IN DECISION MAKING BETWEEN HUMANS AND INTELLIGENT MACHINES

Uzowulu Onyeka Emmanuel
Oe.uzowulu@unizik.edu.ng
&
Prof. Charles Nweke
cc.nweke@unizik.edu.ng

ABSTARCT
The accelerating advancement of Artificial Intelligence (AI) and intelligent machines presents unprecedented challenges to human-centered ethical decision-making. As these systems increasingly participate in critical domains such as healthcare, law, business, and security, the research investigates case studies where intelligent machines assist or substitute human judgment, highlighting both the potential for improved efficiency and the risks of moral displacement, bias, and opacity. The research addresses the question, how should ethical responsibility be shared or distinguished between humans and intelligent machines? Are there ethical frameworks that guide the decision-making processes in contexts where human judgment and machine intelligence intersect? By comparing human ethical reasoning with algorithmic logic, the research emphasizes the limitations of machine-led decision-making in capturing moral nuance, empathy, and contextual sensitivity. In the end, the researchers draw on normative theories, particularly teleological (consequence-based) and deontological (duty-based) ethics to examine how principles of fairness, accountability, autonomy, and the prevention of harm can be applied in hybrid human–machine decision environments.
Keywords: Artificial Intelligence, Ethics, Teleology, Deontology, Machine Ethics

INTRODUCTION
The emergence of intelligent machines capable of autonomous decision-making compels philosophy to revisit long-standing questions concerning agency, morality, and responsibility. This research undertakes a philosophical inquiry into ethical decision-making between humans and intelligent machines, situating the discussion within normative ethical theories. Drawing on **teleological** (consequence-oriented) and **deontological (**duty-based) perspectives, the study critically examines how traditional philosophical frameworks can illuminate the challenges of bias, accountability, autonomy, and moral responsibility in human–machine interaction.

Rather than treating intelligent machines merely as tools, this work interrogates whether and to what extent they can be regarded as moral agents, or whether moral responsibility must remain exclusively human. By analyzing real and hypothetical cases where machine intelligence intersects with human judgment, demonstrates both the possibilities and the limits of applying classical ethical theories to contemporary technological contexts. The central argument advanced is that intelligent machines, while capable of simulating decision-making processes, lack the intentionality and moral autonomy required for genuine ethical agency. Consequently, human beings must remain the ultimate bearers of moral responsibility, while philosophical ethics should guide the design and governance of intelligent systems to ensure they align with fundamental moral principles.

Nath and Sahu (2020) note in the article, *The problem of machine ethics in artificial intelligence*, that the advent of the intelligent robot has occupied a significant position in society over the past decades and has given rise to new issues in society. As we know, the primary aim of artificial intelligence or robotic research is not only to develop advanced programs to solve our problems but also to reproduce mental qualities in machines. The critical claim of artificial intelligence (AI) advocates is that there is no distinction between mind and machines and thus they argue that there are possibilities for machine ethics, just as human ethics. Unlike computer ethics, which has traditionally focused on ethical issues surrounding human use of machines, AI or machine ethics is concerned with the behaviour of machines towards human users and perhaps other machines as well, and the ethicality of these interactions. The ultimate goal of machine ethics, according to the AI scientists, is to create a machine that in itself follows an ideal ethical principle or a set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take (Anderson and Anderson 2007). Thus, machine ethics is tasked with ensuring ethical behaviour of an artificial agent. Although, there are many philosophical issues related to artificial intelligence, but our attempt in this paper is to discuss, first, whether ethics is the sort of thing that can be computed. Second, if we are ascribing mind to machines, it gives rise to ethical issues regarding machines. And if we are not drawing the difference between mind and machines, we are not only redefining specifically human mind but also the society as a whole. Having a mind is, among other things, having the capacity to make voluntary decisions and actions is the main issue. The notion of mind is central to our ethical thinking, and this is because the human mind is self-conscious, and this is a property that machines lack, as yet.

The ultimate objective of machine ethics is to design machines that can independently operate according to an ideal ethical principle or set of principles. In other words, such machines would allow these principles to guide their choices when faced with different possible actions. James Moor (2006) distinguishes between what he calls an *implicit ethical agent* and an *explicit ethical agent*. An implicit ethical agent is a machine programmed to behave ethically or at least to avoid unethical conduct, without directly representing ethical principles. Its ethical boundaries are set by the programmer who embeds these principles during design. By contrast, an explicit ethical agent is capable of reasoning about ethical dilemmas by applying ethical principles directly. It can "represent ethics explicitly and act effectively based on this knowledge." From Moor's perspective, the ultimate aim of machine ethics is to create machines that qualify as explicit ethical agents.

The difficulties confronting researchers in machine ethics can generally be divided into two categories: **philosophical issues** concerning whether ethics can truly be computed, and **technical challenges** arising from the field of AI. The first category raises the fundamental question of whether moral reasoning lends itself to computational processes. One ethical framework that suggests a positive answer is *act utilitarianism*. As a teleological theory, act utilitarianism holds that the moral worth of an action depends entirely on its consequences: the right action is that which produces the greatest net benefit for all those affected, giving equal weight to each individual. A process Jeremy Bentham (2017) describes as a form of *moral arithmetic*. However, before such calculations can be performed, one must define what qualifies as a "good" or "bad" outcome. The most widely recognized variant, *hedonistic act utilitarianism* evaluates actions by the pleasure and pain they generate. Following Bentham's reasoning, such evaluation would require a scale capable of accounting for variables such as intensity and duration of the pleasure or pain experienced by those impacted. Humans, too, would need this information to apply the theory consistently. For AI research, obtaining and structuring such data remains a major challenge, although it is distinct from the problem of actually computing the morally correct choice once the data is available.

In principle, if provided with the necessary inputs, a machine could apply act utilitarianism as reliably as a human. Implementing hedonistic act utilitarianism algorithmically is relatively straightforward: the system would evaluate all possible actions and select the one that maximizes overall net pleasure. This requires input such as the number of individuals affected and, for each, the intensity of the expected pleasure or pain (e.g., on a numerical scale), its duration (e.g., measured in days), and the probability of its occurrence. For each individual, the algorithm multiplies intensity, duration, and probability to arrive at a net utility score, which then informs the final ethical decision.

Brundage (2014) believes that machines may, in fact, hold certain advantages over humans when applying the principles of act utilitarianism. First, while humans often rely on rough estimations rather than precise calculations when weighing outcomes, machines can perform the necessary arithmetic systematically, thereby reducing the likelihood of error. Second, human decision-making is frequently influenced by partiality, favouring oneself, family, or close associates, whereas machines could be designed to operate impartially. This is significant because act utilitarianism was originally developed to inject objectivity into moral decision-making. Third, humans typically overlook many possible alternatives in a given situation, while a machine could be programmed to evaluate a broader range of options. Envisioning a machine advisor that "reasons" like an act utilitarian, it could prompt individuals to consider actions they might otherwise dismiss and to assess the consequences of those actions for all affected parties. Finally, in high-stakes contexts, such as decisions made by national leaders or corporate executives, the potential impact is so vast that calculating the greatest net benefit becomes a demanding task, one that machines, given their processing speed, could accomplish far more efficiently.

From this perspective, machines are capable of following act utilitarianism at least as well as, they are potentially better, while humans provided they are given the necessary information. Nevertheless, the theory itself has long been criticized for failing to align fully with moral intuition. While it offers a promising foundation for programming ethically responsive machines, arguably, it is more reliable than many human judgments which may not represent the most adequate moral framework. Critics contend that act utilitarianism can justify violations of individual rights by sacrificing one person for the greater good, and it often clashes with notions of justice, which are rooted in desert based on past actions rather than future consequences.

Deontological ethics, such as Kant's categorical imperative, addresses this limitation by emphasizing duties, rights, and justice independently of consequences. Yet, this approach is also criticized for disregarding outcomes altogether. Following W. D. Ross (1930), a more balanced ethical framework may be preferable; one that integrates elements of both teleological and deontological theories. Ross's notion of **prima facie duties**, moral obligations that generally ought to be upheld but can be overridden by more pressing duties in specific cases—

captures the complexity of ethical decision-making more effectively than rigid adherence to a single, absolute principle.

This discussion focuses specifically on the process of ethical decision-making, rather than on how a machine might collect and process the information necessary for such decisions. The distinction matters, since access to vast amounts of data or computational ability alone does not ensure ethical behavior. To define what counts as morally acceptable action, we must look to philosophy, particularly ethics. Yet, this is a profound challenge because ethics is not a fully codified or settled discipline; it is still evolving. Indeed, one potential contribution of machine ethics is that it may stimulate progress in ethical theory itself, as machines provide a unique platform for testing the consistent application of moral frameworks.

DECISION MAKING AND ARTIFICIAL INTELLIGENCE

Nowadays, advanced technologies such as robots and artificial intelligences are more integrated to our society and everyday life. Currently, AI is used increasingly in important decision-making. According to Zhang et al. (2023) most AI machines can solve moral dilemmas without humans and act as moral agents whose decisions have consequences for human lives. For example, AI machines are used to help with recruiting processes for matching organs to recipients or for cancer diagnosis. Despite their benefits, the usage of AI in decision-making has raised both moral philosophical and ethical concerns that have been addressed widely. Ferrer et al. (2020) assert that AI systems do produce discriminative and biased decisions, for example, based on gender and ethnic background.

The questions related to AI, its usage and development are connected with societal power structures, culture, and politics but these are often discussed only as technical details. However, it is important to consider who has the power over these technologies and decisions that can be made with them, more precisely how they are used and by whom. That is why it is also important to understand how people perceive AI-made decisions that might be biased. It is widely documented that decisions made by robots or AIs are viewed differently than those made by humans. For instance, Bigman et al. (2023) observes that people express lower levels of moral outrage towards AIs making discriminative decisions compared to humans. Perhaps soon, we will have AIs that can operate more independently and whom we may even blame for breaking norms. At the same time these machines must be able to solve complex moral dilemmas. A tangible examples of AI technologies that can also act as decision makers are self-driving cars that are able to navigate traffic autonomously. David C. Vladeck (2014), in *Machines without principals: liability rules and artificial intelligence,* notes that machines at this level usually face situations where they need to make sacrificial decisions in order to prevent fatal accidents. To develop the ethical guidelines for how machines should act or intervene in different situations, many moral and social philosophers have drawn attention to ordinary people's views of decision-making algorithms.

Consequently, the reality as to whether people judge machine and human made moral decisions differently, Micheal Laakasuo et al. (2021) believe that earlier studies have shown that people perceive AI and human-made decisions differently, even if circumstances and consequences of decisions are the same. To portray this point, there is need to conduct an experiment where people's moral judgments of human and AI decision-makers will be compared in a situation when an agent conducts a norm-conflicting decision that has harmful consequences for a moral patient. However, it is important we look at some moral dilemmas that is associated with decision making in AI technologies.

DEONTOLOGY AND UTILITARIANISM IN THE TROLLEY PROBLEM'S MORAL DILEMMAS

Technological development has increased the number of interactions people have with intelligent machines. This has also created a whole new category of moral philosophy focused on intelligent machines. Traditionally discussion on AI remains focused on people, supernatural beliefs, and animals, moral philosophy but here, it seeks answers to how ordinary people react to machines that make decisions. At the same time, it investigates how these perspectives should be used, for example, in developing legislation or designing robots that better meet expectations of moral competence. While the ethics of AI raises new questions, tools to study human morality have remained structurally similar but developed further nuances and are used in a new context. For instance, classic trolley problem dilemmas is widely used in moral philosophy to investigate, for example, moral intuitions and moral judgments.

Today, researchers are applying these to study human judgments on self-driving cars and robotics. The trolley problem thought experiment, directs us to consider the right options in a situation where multiple people can be saved if one is sacrificed. Garry Young (2025), explains the dilemma by stating that;

It involves when people are standing on the railway tracks and going to get crashed by a broken trolley. However, they can be saved if the trolley is redirected to the other track by pulling a switch. By redirecting the trolley, only

one innocent person would be killed, but the other five would be spared. There are also other adaptations of the trolley problem such as the footbridge and lifeboat dilemmas. In the footbridge dilemma, five people are in danger and going to get hit by a trolley. However, this time, they can be saved if a bystander is pushed to the tracks. This will stop the trolley, and only one person will be sacrificed. Finally, in the lifeboat dilemma, the decision-maker must decide whether to push an injured person out of the overcrowded lifeboat to save the other passengers. Otherwise, every passenger will drown as the lifeboat cannot carry the weight of all passengers.

At the same time, these thought experiments pit against two normative moral philosophical theories that can be seen as the principles for guiding actions: deontology and utilitarianism. Deontology as a moral theory explains how some actions are good or forbidden. These rules can be applied universally. In a deontological sense, good consequences cannot justify the harm that actions could bring. From the deontological perspective it is not allowed to use someone just as a means to achieve a desired outcome (e.g., pushing one from a bridge to prevent an accident). Utilitarianism has a fundamentally different approach to defining the goodness of actions. Instead of focusing on the rightness of actions, the focus is on maximizing good consequences. The more people benefit from these consequences, the better. Usually this means maximizing welfare or happiness. However, in some instances, utilitarianism demands that innocent people may die if their death benefits others.

Accordingly, Sommaggio and Marchiori (2020) in their work, *Moral dilemmas in the AI era: A new approach* argue that deontology and utilitarianism have both different stances on solving trolley problem-like dilemmas. In a utilitarian sense, it would be a permitted option to take the action as it would save more people while the deontologist would make contradicting decisions and refuse to harm others. Kohlberg (1996) was the first to apply moral dilemmas to investigate human morality to explain how moral reasoning, that is, the reasons people give when solving moral conflict, develops. Since then, moral philosophy research in terms of moral decision-making has developed further.

Greene et al. (2004) measure brain activity when participants were presented with moral dilemmas and found that the dilemmas that included different levels of personal harm were solved oppositely. Participants found it more acceptable to make utilitarian decisions when solving the trolley problem dilemma's switch version. Turning a switch to change the course of the trolley to conduct a utilitarian decision included less personal harm. When the switch was used as a tool, it made the act more impersonal and indirect. In contrast, participants felt it was harder to make the utilitarian choice in the footbridge dilemma, even though the outcomes are similar in both dilemmas: sacrificing multiple people is inevitable. However, the footbridge dilemma included more personal and direct harm, as the agent should actively push the other person off the bridge without any instruments, such as the switch in the trolley dilemma. Finally, it is important to note that the results do not justify the actions. Instead, the results highlight the role of cognitive and emotional processing involved in solving moral dilemmas and how they differ across dilemmas. Interestingly, the different kinds of dilemmas engaged distinct brain regions. Dilemmas which included direct personal harm showed greater activity in brain areas connected to emotional processing, which affected the moral judgments.

Recently, moral philosophy has shifted also to consider the role of intuitions when making moral judgments. Kesebir and Haidt's (2010) social intuitionist model proposes that people make moral judgments quickly and automatically when thinking about or observing norm-breaking actions. If needed, these intuitions are followed by moral reasoning, which may correct the initial moral response when more information becomes available. Indeed, Kesebir and Haidt (2010) have emphasized the social purpose of moral thinking. Relevant information to support moral judgments is gathered in a social context, for example, by gossiping. Moreover, morality has an important role in community building as it keeps people together. Moral communities share norms that guide the behaviour of group members and support, for example, peace and cooperation within the group. Finally, they present that "morality is about more than harm and fairness". By this they mean that some of the moral intuitions are also linked with, for example, loyalty, authority, or respect or obedience. Finally, although moral scenarios like the trolley problem are hypothetical, these dilemmas and questions related to them have become relevant as technology has advanced, for example, due to the adoption of autonomous vehicles. As a result, autonomous vehicles could face, for example, a low-probability situation where they need to make a moral decision about whom to spare in an accident. The moral psychology of AI aims to understand how humans judge machine-made moral decisions by using traditional methods and applications to study new emerging technologies.

MORAL DECISIONS MADE BY HUMANS VS INTELLIGENT MACHINES

Over the years, social and moral philosophical experiments aim to study human responses and perceptions of robots in solving moral dilemmas have expanded. According to Moser, Den Hond, and Lindebaum (2022), studies seem to consider intelligent machines as explicit moral agents. The first experiment in the field by Malle (2016) provides an indication that there are differences in people's moral expectations towards robots and humans. She

observed that people rather wanted human decision-makers to make deontological choices in situations where multiple persons can be saved if one is sacrificed. Robots were expected to make the opposite choice, and their decision to sacrifice one to save others in danger was considered less wrong. Additionally, people were more likely to blame robots when they made deontological choices. She also brought up that moral judgments on how humans influenced the way robots were judged, perhaps setting standards for comparison. Finally, they concluded that robots must be built on people's preferences, whether it means making utilitarian choices or being an active part in risky situations. Additionally, to be trusted, robots should be able to explain their moral choices even if they do not always meet moral expectations. As such, the study showed that moral norms can also be applied when evaluating robot decision-makers in terms of blame and wrongness judgements.

Further research according to Dhirani et al. (2023) have expanded quickly and developed more complex moral dilemmas to address the difficulties in developing moral rules for emerging technologies. The scenarios have shifted from theoretical trolley-problem-like dilemmas to more realistic alternatives to address current and future AI and robotics development. Bello and Bringsjord (2013), state that one of the most extensive experiments in the field is MIT's Moral Machine experiment, which gathered almost 40 million responses for moral dilemmas from 233 countries. According to them,

…the experiment was conducted as an online game in which autonomous vehicles (AV) would face unavoidable sacrificial situations. Participants were presented with dilemmas to solve and choose the targets they would prefer to be sacrificed. Dilemmas included nine different factors from which participants could choose. For example, the AV could either stay on course or deviate, causing either death of the pedestrians or passengers of the AV. In addition, scenarios varied between genders (males vs. females), socioeconomic status (low vs. high SES), age (young vs. old), and utilitarianism (saving more vs. fewer lives). Finally, some scenarios also included characteristics about fitness, the lawfulness of pedestrians walking the crossroad, and species (humans vs. pets). The Moral Machine experiment suggested that people strongly prefer saving more lives, saving humans over pets, and saving youngsters. Finally, it is important to notice that the participants judged only machine-made decisions as there was no comparison to human drivers.

In extending the argument, Edmond Awad (2017) in the work, *Moral machines: perception of moral judgment made by machines*, conducted three experiments to measure moral wrongness and moral blame judgments in a military context. He compared people's judgments on autonomous drones, AI agents, and human drone pilots. In the experimental task, an agent was described to decide whether to cancel or launch a missile strike to prevent a terrorist attack. However, preventing the attack would harm an innocent child. Artificial agents' decisions to launch the missile were considered more wrong than cancelling. Results were reversed for the human pilot as they were judged more and received more blame for the decision to cancel the launch. Malle (2016), named this phenomenon human-machine asymmetry that was caused by different expectations. The human pilot was seen as a part of a military command chain where commands are strictly followed. The human pilot's decision to launch the missile received less blame because the agent was performing their duty, whereas the decision to cancel was seen as breaking the command chain, which received more blame. Blame judgments were modified by different assumptions of the agent's role in society. According to Malle (2016), the results indicate that human pilot was seen to operating within social structures and following moral justifications. However, the actions of artificial agents were not considered as similarly embedded part of society (only a small portion of participants saw the artificial agents this way). Interestingly, many participants considered AI agents as suitable recipients of moral blame compared to autonomous drones, which were seen more as machines.

Form the above, one can observe that we utilize our moral cognition and concepts when judging artificial agents' decisions. The differences in blame judgments show that people explain AI agents' actions differently as those are not considered to be part of society. Furlough, Stokes, and Gillan (2021) agree because they examined blame judgments attributed to robots in military, warehouse, and surgical environments. Their studies indicated that robots received less blame than humans for mistakes but still more than the environmental factors causing the failure in the task. However, robots that were described to be fully autonomous received more blame than robots without the same description. Fully autonomous robots were blamed almost as much as humans, whereas non-autonomous robots received the same level of blame as the environmental factors. One can notice that human decision-makers can be seen as more independent decision-makers and in contrast, robot decision-makers are perceived as more computational without the same level of autonomy as humans (Furlough et al (2021).

Human attitudes toward machine-made decisions have also been studied in a care work context. Laakasuo et al. (2021), examined people's views in situations in which a robot or a human nurse needs to forcefully medicate a patient. In the extensive research included an anthropological field study and four experimental studies, was conducted to understand how seniors living in residential care homes perceive using robots in care work. It showed that the participants were cautious about the care robots. Participant's concerns were related to, for example, fear

of losing autonomy and lack of trust. The care robots were perceived as incapable of showing empathy and seen as cold. The anthropological study inspired later experimental studies. The experiments showed that robot nurses were judged more than their human colleagues when acting against patients' autonomy by forcefully medicating the patient. In situations where the patients' autonomy was prioritized by ignoring the doctor's orders, the human and robot nurses' actions received almost the same level of approval. These findings are consistent where AI agents' violations of human orders were preferred over obedience. To conclude, it seems that involuntary medication decisions are more justified when they are made by a human than by a robot. However, both robot and human nurses are expected to act autonomously when resisting doctor's orders and valuing patient's right to autonomy.

Having compared moral judgments toward rescue robots and human lifeguards, it is recorded that such experiments manipulated rescue agents (human vs. robot) and robots' mental (automata vs. human-like mind) and physical features (floater drone vs. android) (J. Sundvall et al. (2023). Additionally, the properties of the rescued party (motorboater vs. fisher) were manipulated in terms of culpability and utilitarianism. This demonstrated that laypeople prefer saving the innocent over the ones who caused the accident. Interestingly, a robot saving the two culpable parties instead of one innocent was judged more harshly than a human taking the same action. Generally, people viewed robot's decisions as bad when they violated moral standards. However, when robots managed to meet the standards, their decisions were evaluated similarly to humans.

Some of the previous research experiments by Marcondes, Francisco (2012), in his article, *The Moral Dilemma of Computing Moral Dilemmas*, shows that AI machines are expected to have "higher moral standards" than humans. However, it is not clear what the higher standards are a priori. Perhaps, deduced post hoc, people might expect robots to be more computational, and when these standards are failed, they are judged worse than their human counterparts. Additionally, moral standards seem to vary in different situations; for example, Longoni and Cian (2022) notice people preferred the agent making a utilitarian decision when both parties to be saved were culpable, even though overall, people preferred saving the innocent one. This contradicts some earlier findings where people have shown a robust utilitarian preference. People were less morally outraged toward AI than human recruiters who made discriminative decisions. However, when the algorithm was described to have the same capabilities as human recruiters (i.e., the algorithm was anthropomorphized), the level of moral outrage was higher. According to Bigman et al. (2023), people might not see algorithms as having similar prejudiced motivations to humans. Therefore, participants were less morally outraged about algorithmic discrimination.

Stuart and Kneer (2021) have observed a similar pattern in AI or robot blame but offered a different explanation. They noticed that people might avoid blaming robots or AIs for bad outcomes. Perhaps participants understood that at the end of the day, the outcomes that artificial agents produce are programmed by humans, so they are the ones who should be held morally responsible, not just the robot or AI. Blaming the robot would help humans get away with moral responsibility. In other words, there is a lot of diversity in how ordinary people view machine-made moral decisions. Judgments seem to vary depending on, for example, the context and the perceived level of the agent's autonomy. J.W. Sundvall et al. (2016), notice that people are allowed to make more bad decisions than robots. People might see robots as more rational than humans, and when the decision conflicts with expectations they are judged more harshly. Sometimes, it seems that robots are expected to disobey when needed, but humans are left more space to decide whether to obey or disobey. It seems that robots are easily blamed, whatever decision they take. However, sometimes people might avoid blaming robots as they recognize that instead they should blame humans. Finally, it seems that people expect robots and humans to act based on different norms, but we cannot foresee what those norms are in which situation.

CONCLUSION

In the light of research on artificial intelligence, bias can be defined as a difference in technical performance that benefits or creates disadvantages for a certain group. Biases infiltrate the systems, for example, through insufficient data that are used to train the algorithm or by people who are building it. Machine ethics has emerged as an important and promising branch of artificial intelligence, with the primary aim of developing systems that function as explicit ethical agents. Preliminary experiments in restricted domains suggest that embedding an ethical dimension within machines is possible. However, enabling such systems to operate autonomously in real-world contexts remains a significant challenge. Researchers must address complex issues such as how to represent and formalize ethical principles, how to incorporate them into decision-making procedures, how to make ethical judgments under uncertainty, how to explain such decisions, and how to evaluate ethically guided systems.

A central requirement in this field is sustained dialogue between ethicists and AI researchers. Such collaboration offers mutual benefits: ethicists may refine or even discover fundamental ethical principles, while AI researchers may strengthen public trust in the possibility of ethical machines, thereby securing support for the continued

development of autonomous systems aimed at improving human life. Expert systems decision-based programs that aim to reproduce expert-level reasoning illustrate both the potential and the limitations of machine ethics. Although they can replicate aspects of human decision-making, most business and organizational choices are based on incomplete or ambiguous information, making full automation difficult. Consequently, the notion of fully autonomous "artificial ethics" remains closer to science fiction than reality. Expert systems operate through symbolic representations of the world, relying on information, inferential rules, and decision criteria, all of which are fallible. The lack of human intelligence, emotional depth, and values in AI machines introduces risks of bias and raises ethical concerns when entrusted with moral decisions.

Therefore, decision-making should not be delegated entirely to expert systems. These technologies function best when assisting with routine, well-structured tasks, thereby freeing human managers to focus on more complex moral and strategic judgments. Managers retain both legal and ethical responsibility for outcomes, and they must critically evaluate system-generated advice rather than accept it uncritically. Since faulty recommendations may appear sound but cause long-term harm, mechanisms such as software standards, regular inspections, and accountability measures (e.g., licensed engineers certifying software) could strengthen trustworthiness.

Lastly, system designers should emphasize the balance between "high-tech" efficiency and "high-touch" human judgment. Instead of replacing human decision-makers, machines should enhance their capacity for accurate analysis, faster information gathering, and more reliable consultation. This requires prioritizing the user interface, which highlights the role of the human operator rather than the machine itself. By doing so, expert systems and human expertise can work together synergistically, ensuring that technology supports rather than supplants ethical responsibility.

REFERENCES

Anderson, Michael, and Susan Leigh Anderson. 2007. "Machine ethics: Creating an ethical intelligent agent." *AI magazine* 28 (4): 15-15.

Awad, Edmond. 2017. "Moral machines: perception of moral judgment made by machines." Massachusetts Institute of Technology.

Bello, Paul, and Selmer Bringsjord. 2013. "On how to build a moral machine." *Topoi* 32 (2): 251-266.

Bentham, Jeremy. 2017. *The Correspondence of Jeremy Bentham, Volume 3: January 1781 to October 1788*. UCL Press.

Bigman, Yochanan E, Desman Wilson, Mads N Arnestad, Adam Waytz, and Kurt Gray. 2023. "Algorithmic discrimination causes less moral outrage than human discrimination." *Journal of Experimental Psychology: General* 152 (1): 4.

Brundage, Miles. 2014. "Limitations and risks of machine ethics." *Journal of Experimental & Theoretical Artificial Intelligence* 26 (3): 355-372.

Dhirani, Lubna Luxmi, Noorain Mukhtiar, Bhawani Shankar Chowdhry, and Thomas Newe. 2023. "Ethical dilemmas and privacy issues in emerging technologies: A review." *Sensors* 23 (3): 1151.

Ferrer, Xavier, Tom Van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. 2020. "Bias and discrimination in AI: a cross-disciplinary perspective." *arXiv preprint arXiv:2008.07309*.

Furlough, Caleb, Thomas Stokes, and Douglas J Gillan. 2021. "Attributing blame to robots: I. The influence of robot autonomy." *Human factors* 63 (4): 592-602.

Greene, Joshua D, Leigh E Nystrom, Andrew D Engell, John M Darley, and Jonathan D Cohen. 2004. "The neural bases of cognitive conflict and control in moral judgment." *Neuron* 44 (2): 389-400.

Kesebir, Selin, and Jonathan Haidt. 2010. "Morality (in Handbook of social psychology)." *Handbook of social psychology, 5th Ed., S. Fiske, D. Gilbert, & G. Lindzey, eds., Forthcoming*.

Kohlberg, Lawrence. 1996. "Moral reasoning." *Educating the democratic mind*: 201-221.

Laakasuo, Michael, Volo Herzon, Silva Perander, Marianna Drosinou, Jukka Sundvall, Jussi Palomäki, and Aku Visala. 2021. "Socio-cognitive biases in folk AI ethics and risk discourse." *AI and Ethics* 1 (4): 593-610.

Longoni, Chiara, and Luca Cian. 2022. "Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect." *Journal of Marketing* 86 (1): 91-108.

Malle, Bertram F. 2016. "Integrating robot ethics and machine morality: the study and design of moral competence in robots." *Ethics and Information Technology* 18 (4): 243-256.

Moor, James H. 2006. "The nature, importance, and difficulty of machine ethics." *IEEE intelligent systems* 21 (4): 18-21.

Moser, Christine, Frank Den Hond, and Dirk Lindebaum. 2022. "Morality in the age of artificially intelligent algorithms." *Academy of Management Learning & Education* 21 (1): 139-155.

Nath, Rajakishore, and Vineet Sahu. 2020. "The problem of machine ethics in artificial intelligence." *AI & society* 35 (1): 103-111.

Sommaggio, Paolo, and Samuela Marchiori. 2020. "Moral dilemmas in the AI era: A new approach." *Journal of Ethics and Legal Technologies* 2 (JELT-Volume 2 Issue 1): 89-102.

Stuart, Michael T, and Markus Kneer. 2021. "Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1-27.

Sundvall, Jonathan Widén, Peter Nymberg, Carl Wikberg, Anna Moberg, and Ronny Gunnarsson. "Exploring Primary Care Patients' Perspectives on Artificial Intelligence: A systematic literature review and qualitative meta-synthesis."

Sundvall, Jukka, Marianna Drosinou, Ivar Hannikainen, Kaisa Elovaara, Juho Halonen, Volo Herzon, Robin Kopecký, Michaela Jirout Košová, Mika Koverola, and Anton Kunnari. 2023. "Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas." *European Journal of Social Psychology* 53 (4): 779-804.

Vladeck, David C. 2014. "Machines without principals: liability rules and artificial intelligence." *Wash. L. Rev.* 89: 117.

Young, Garry. 2025. "Using the classic trolley problem to teach AI students and researchers about their role as moral agents, and why they should be subject to moral scrutiny." *AI and Ethics* 5 (2): 1877-1883.

Zhang, Yuyan, Jiahua Wu, Feng Yu, and Liying Xu. 2023. "Moral judgments of human vs. AI agents in moral dilemmas." *Behavioral Sciences* 13 (2): 181.