Ethical thinking, AI vs Humans

Rares BRAILEANU
Academia de Studii Economice Bucuresti (FABIZ), Bucharest, Romania
braileanurares24@stud.ase.ro
Darius NITU
Academia de Studii Economice Bucuresti (FABIZ), Bucharest, Romania
nitudarius24@stud.ase.ro

Abstract. *The rapid development of artificial intelligence has raised important questions about ethical thinking and whether machines can truly replicate human moral judgment. This paper examines the differences between ethical reasoning in AI systems and in humans, focusing on decision-making, empathy, accountability, and contextual understanding. While AI can process large amounts of data quickly and apply programmed ethical frameworks consistently, it does not possess consciousness, emotions, or lived experience. Human ethical thinking, by contrast, is shaped by empathy, social values, cultural background, and the ability to reflect on complex moral dilemmas beyond fixed rules. This study combines a small-scale survey of human respondents with a structured comparison of answers provided by six AI systems to a set of moral dilemmas. The findings show both convergence and divergence: humans and AI agreed on several classic prohibitions, such as rejecting the sacrifice of a healthy patient, but differed more strongly when dilemmas involved personal sacrifice, loyalty, and responsibility. The paper argues that AI should not be viewed as a replacement for human ethical judgment, but rather as a tool that can support ethical decision-making in areas such as healthcare, law, and business. Ultimately, the comparison between AI and humans shows that ethical thinking is not only a matter of logic, but also of values, compassion, and responsibility. The study concludes that the most effective approach is human-AI collaboration, where AI assists analysis while humans retain final ethical authority.*

Keywords: AI ethics, moral dilemmas, human judgment, machine ethics, accountability, human oversight.

Introduction

Artificial intelligence is now used to recommend medical treatments, detect fraud, classify legal documents, evaluate credit risk, and assist with hiring, education, and public administration. As these systems increasingly move from narrow prediction tasks to socially significant decisions, a central question emerges: can artificial intelligence think ethically in a way that is comparable to human beings? The present paper addresses this question through a direct comparison between human answers and AI-generated answers to a set of moral dilemmas. Rather than assuming that speed, consistency, or computational power automatically translate into moral competence, the paper investigates what ethical thinking actually looks like when both humans and machines are asked to choose between competing values.

The topic is important for at least three reasons. First, modern AI systems are already involved in contexts where ethical consequences are unavoidable. Even when the final decision remains formally human, automated outputs can frame the range of acceptable choices and influence who receives care, attention, resources, or sanctions. Second, public debate often oscillates between exaggerated optimism and exaggerated fear. Some narratives suggest that advanced AI will soon outperform humans in every aspect of reasoning, including moral reasoning, whereas others assume that machine decision-making is inherently dangerous and should never be trusted. A more useful approach is empirical: observing where AI responses resemble human moral intuitions, where they diverge, and what that divergence reveals about the

DOI:

nature of ethical judgment. Third, the question is not merely technical. It concerns accountability, legitimacy, social trust, and the limits of delegating moral authority to computational systems.

The literature already provides several starting points for this comparison. Research on moral cognition has shown that human judgment is not reducible to abstract logic alone. Classic trolley-style dilemmas reveal stable patterns in the way people distinguish between indirect harm and direct personal violence, which suggests that emotional responses and contextual framing are deeply involved in ethical reasoning (Greene, 2016). Research on AI ethics, by contrast, has focused on principles such as fairness, transparency, accountability, privacy, and human oversight, emphasizing that systems may produce harmful outcomes even when they appear technically efficient (Floridi et al., 2018; Fjeld et al., 2020; Mittelstadt, 2020). Together, these strands of literature imply that the comparison between AI and humans should not be limited to asking whether answers match. It should also ask why they match, why they differ, and what forms of reasoning those patterns appear to express.

The present study uses original survey data collected through Google Forms from human respondents and compares them with answers produced by six AI systems: ChatGPT, Gemini, Duck.ai, Copilot, Perplexity, and DeepAI. The comparison focuses on ten shared moral dilemmas, including the trolley problem, lying to protect a friend, organ sacrifice, self-driving car priorities, distributive justice, self-sacrifice, punishing an innocent person, loyalty to a friend who committed fraud, and the value of life dilemma. The human questionnaire also included a final perception item asking whether AI can make ethical decisions comparable to humans. This item does not have an AI equivalent and is therefore treated separately as an attitudinal indicator rather than as a comparative dilemma.

The main research question is the following: how similar are AI and human responses when both are confronted with the same moral dilemmas, and what do similarities or differences reveal about the limits of machine ethics? The study advances three working hypotheses. First, AI systems will appear more outcome-oriented and more willing than humans to endorse decisions that maximize aggregate benefit, especially when emotional self-cost is involved. Second, human respondents will show stronger reluctance toward direct harm, betrayal, and the sacrifice of personal relationships. Third, the comparison will suggest that even when AI outputs resemble human answers on the surface, the resemblance does not imply equivalent moral understanding, because machine outputs are produced without lived experience, emotional accountability, or personal vulnerability.

The contribution of the paper is modest in scale but relevant in focus. It does not claim to settle the philosophical question of whether machines can ever become moral agents. Instead, it provides a small empirical comparison that connects everyday intuitions about ethics with broader debates on machine decision-making. By combining descriptive statistics, visual comparison, and literature-based interpretation, the paper shows that ethical thinking cannot be understood as rule application alone. It is also a matter of contextual sensitivity, responsibility, and the social meaning of decisions. For that reason, the paper argues that the most defensible model is not AI replacing humans in moral judgment, but AI supporting human deliberation under clear human responsibility.

Literature review

Human moral judgment and machine ethics

Ethical thinking is notoriously difficult to define because it includes more than one dimension. In philosophy, ethics concerns judgments about right and wrong, duties and consequences, justice and care, rules and virtues. In psychology, it also involves cognitive mechanisms, emotional responses, social learning, and identity. In practical governance, ethics concerns accountability for action, the legitimacy of procedures, and the protection of persons affected by decisions. For the purposes of this paper, ethical thinking is understood as the capacity to evaluate competing courses of action in light of moral values, foreseeable harms, and responsibilities toward others. This definition is broad enough to accommodate both principle-based and context-sensitive reasoning, while still allowing comparison between human respondents and AI outputs.

Human moral judgment has often been studied through dilemmas that create tension between consequentialist and deontological intuitions. Greene (2016) argues that trolley-type cases are useful because they reveal a recurring pattern: many people are willing to redirect a threat by pulling a lever, yet far fewer accept physically pushing a person to his death in order to save a larger number. The contrast suggests that human moral reasoning combines outcome evaluation with strong intuitive aversion to direct personal harm. This dual structure matters for the present study because several of the selected dilemmas reproduce exactly this tension. The question is not only whether respondents save more lives, but how they react when saving lives requires direct agency, betrayal, or self-sacrifice.

The relevance of such findings extends beyond philosophy classrooms. Moral dilemmas are simplified scenarios, but they illuminate deeper features of judgment: the role of emotional distance, the importance of perceived intent, the distinction between killing and letting die, and the influence of personal relationship. Human beings do not merely calculate totals. They interpret the meaning of action within a social and personal context. Lying to protect a friend, refusing to punish an innocent person for public stability, or declining to report a close friend for a minor fraud may all appear inconsistent from a purely rule-based perspective, yet they express values such as loyalty, care, trust, and the rejection of instrumentalizing persons. In this sense, variability in human responses should not automatically be viewed as irrational noise; it may also indicate the presence of competing moral commitments.

AI ethics, accountability, and oversight

The literature on AI ethics approaches the problem from a different angle. Instead of asking how emotions and intuitions shape moral judgment, it asks how computational systems should be designed, constrained, explained, and governed when their outputs affect people. Floridi et al. (2018) synthesize a widely cited framework built around beneficence, non-maleficence, autonomy, justice, and explicability. These principles are not a recipe for machine conscience. Rather, they are governance conditions intended to keep AI aligned with human values and social expectations. Their importance lies in showing that even well-performing systems require ethical architecture around them. Accuracy alone does not guarantee justice, and efficiency alone does not guarantee legitimacy.

A similar point appears in work mapping convergence across major AI principles documents. Fjeld et al. (2020) identify transparency, justice and fairness, non-maleficence, responsibility, privacy, and accountability as recurring themes across institutional frameworks. The convergence is significant because it shows broad agreement that the ethics of AI is not reducible to technical optimization. Systems must be understandable enough to be scrutinized, fair enough to avoid systematic harm, and accountable enough that responsibility does not disappear into the model. The strong recurrence of accountability is especially relevant for this paper, because one of the persistent differences between humans and AI lies in the fact that humans can be blamed, can justify themselves, and can carry moral burden in a way machines cannot.

Doshi-Velez and Kim (2017) place explanation at the center of AI accountability. Their argument is particularly important in high-stakes settings such as law, medicine, and public administration, where decisions cannot simply be accepted because an algorithm produced them. Ethical acceptability depends not only on the outcome but also on whether stakeholders can understand the basis of the decision and contest it if necessary. This concern immediately distinguishes machine output from human moral agency. A human decision-maker may still be biased or mistaken, but in principle that person can be asked to justify the decision, reflect on criticism, revise the judgment, and assume responsibility. AI systems, by contrast, do not take responsibility in the moral sense; responsibility remains with the people and institutions that design, deploy, or rely on them.

Mittelstadt (2020) adds another important layer by warning that the ethics of AI itself is not a settled or purely technical exercise. Ethical analysis of AI faces conceptual ambiguities, implementation limits, prediction problems, and tensions between narrow problem-solving and broader systemic critique. This observation matters because popular discourse sometimes assumes that ethical frameworks can simply be encoded into machines. In practice, however, moral concepts are contested, context-dependent, and often difficult to formalize without loss. A model may be trained to prefer one principle over another, but deciding which principle should dominate in the first place is a human normative task. Consequently, claims that AI can "solve" ethics often hide the prior human labor of defining labels, designing prompts, choosing datasets, and interpreting results.

Public preferences, cultural variation, and the research gap

Empirical research on machine ethics also shows that public preferences are neither uniform nor universal. The Moral Machine experiment gathered millions of decisions across countries and found broad patterns as well as substantial cultural variation in moral priorities related to autonomous vehicles (Awad et al., 2018). The authors show that preferences can differ on issues such as sparing the young, obeying traffic rules, or minimizing the total number of deaths. For the present paper, the significance of this study is twofold. First, it demonstrates that even among humans there is no single uncontested moral algorithm. Second, it suggests that programming AI for ethical behavior cannot be separated from social context, because different communities may rank values differently. This makes the search for universal machine ethics more complex than simply matching one average response profile.

Recent legal and governance debates further reinforce the importance of keeping humans involved in consequential decisions. Work emerging from Oxford and Harvard highlights the risks of assigning apparent neutrality or authority to automated decision systems in domains where rights, dignity, and interpretation matter. Even when AI offers useful support, human oversight remains crucial because social decisions are embedded in norms, institutions, and relationships that cannot be fully represented through statistical pattern recognition alone (Doshi-Velez & Kim, 2017; Mittelstadt, 2020). This does not mean AI has no ethical value. On the contrary, well-designed systems can improve consistency, reveal hidden patterns, and reduce certain forms of arbitrariness. But these benefits are compatible with a supporting role rather than full delegation of moral authority.

The literature therefore reveals a gap that motivates the present study. On the one hand, human moral psychology explains why ethical judgment is often contextual, emotional, and sometimes internally conflicted. On the other hand, AI ethics scholarship explains why accountability, transparency, and value alignment are necessary when machine outputs shape decisions. What is less visible in everyday discussion is the direct side-by-side comparison between ordinary human respondents and multiple publicly accessible AI systems confronted with the same dilemmas. A small comparative design cannot capture the whole complexity of either domain, but it can expose meaningful contrasts. It can show where AI mirrors familiar moral patterns, where it departs from them, and how those differences support the broader argument that ethical thinking involves more than consistent rule application.

Methodology

The research design is comparative and exploratory. It compares a small human sample with a small but diverse set of AI systems by asking both to respond to equivalent moral dilemmas. The goal is not to test population-level representativeness, but to identify structured patterns of similarity and difference between human respondents and AI outputs. The study combines descriptive quantitative analysis with qualitative interpretation informed by the literature reviewed above.

Human data were collected through Google Forms. The questionnaire included demographic items on age, gender, field of study or occupation, and self-assessed familiarity with AI, followed by eleven substantive items. Ten of these items corresponded to moral dilemmas that were also presented to AI systems. The eleventh substantive item asked whether respondents believed artificial intelligence can make ethical decisions comparable to humans. The final dataset contained twelve human responses. One participant did not answer the first dilemma, which is why the valid number of responses for that item is eleven rather than twelve.

The AI dataset consists of answers from six systems: ChatGPT, Gemini, Duck.ai, Copilot, Perplexity, and DeepAI. Each system received the same prompt structure. The instruction explicitly asked for only the selected answer, without explanation, so that the outputs would be as comparable as possible. The AI prompt contained twelve items in total. Ten of them matched the human survey dilemmas used in this paper. Two additional AI-only items concerned privacy versus security and responsibility for AI-caused harm. Because no equivalent human responses

were collected for these two items, they are discussed only briefly as contextual evidence and are not included in the direct human-AI comparison tables or figures.

The ten shared dilemmas can be grouped into four thematic categories. The first category concerns sacrificial harm: the trolley problem, the bridge variant, and the organ transplant case. The second concerns truth, justice, and institutional legitimacy: lying to protect a friend, convicting an innocent person to prevent riots, and reporting a friend who committed a small financial fraud. The third concerns distributive and protective choices: the self-driving car dilemma, the donation dilemma, and the child versus two adults scenario. The fourth concerns personal cost: whether the respondent would lose ten years of life to save ten strangers. This thematic grouping supports the later discussion by showing that not all differences between humans and AI have the same moral character.

The analysis proceeded in three stages. First, the human dataset was cleaned and recoded so that answer options were standardized. For example, spelling variants such as “the passager” were recoded to “Passenger”, and the final value-of-life item was recoded from “A/B” into “Child/Two adults”. Second, frequency counts and percentages were calculated separately for human respondents and AI systems. Third, the results were visualized through bar charts and summarized in tables embedded in the final paper. Calculations and chart generation were performed programmatically in Python. The Word document was then formatted according to the requirements of the FABIZ Students’ Scientific Session template.

Because the datasets are small, no inferential statistical claims are made. The analysis remains descriptive. Percentages are used to make comparison easier, but the paper does not claim statistical significance. This limitation is especially important for the AI sample, where each model is represented by a single response set. AI outputs can vary depending on prompt phrasing, model version, or hidden system settings; therefore, the study treats each AI answer as an observed output in a specific prompting context rather than as a definitive statement about the model as a whole.

Despite these limitations, the design has several strengths. The dilemmas are simple and intuitive, which reduces ambiguity for respondents. The use of multiple AI systems makes the machine side of the comparison less dependent on a single model. The inclusion of both yes/no dilemmas and multi-option dilemmas allows the analysis to capture not only prohibitions and permissions but also priority structures. Finally, because the study compares actual answers rather than abstract claims, it offers a concrete basis for discussing whether similarity in outputs should be interpreted as similarity in ethical thinking. This distinction is central to the argument of the paper.

Results and discussions

The human sample is small but internally coherent enough to support descriptive comparison. As shown in Table 1, respondents were relatively young, with ages ranging from 19 to 52 and a concentration around ages 19–21. Seven respondents identified as male and five as female. Self-rated familiarity with AI was moderate to high: half of the participants selected the maximum familiarity level of 5, while the rest selected 3 or 4. Figure 1 illustrates this pattern. The sample

therefore represents participants who are not specialists in AI ethics but are sufficiently familiar with the topic to provide informed intuitive judgments.

The first notable result is that humans and AI agree strongly on some baseline prohibitions. In the standard trolley problem, 72.7% of human respondents and all six AI systems chose to pull the lever, which suggests a shared willingness to accept indirect intervention when it saves a greater number of people. Yet this willingness drops sharply in the bridge variant: only 33.3% of humans and one of the six AI systems accepted pushing the person. This contrast reproduces the classic distinction identified in moral psychology between redirecting harm and using a person as a direct means to an end (Greene, 2016). In this area, AI output does not appear radically alien. Most of the models generated the same deontological hesitation that many humans show when harm becomes physically personal.

A similar pattern appears in the organ transplant dilemma. Human respondents rejected sacrificing the healthy patient by a margin of 91.7%, and five of the six AI systems did the same. Even in a highly outcome-oriented scenario, both groups largely refused to instrumentalize an innocent individual. The convergence is important because it shows that AI systems, at least under this prompt, are not mechanically utilitarian in every case. Their answers can reflect widely shared prohibitions against killing one person as a means of saving several others. However, surface convergence should not be overinterpreted. Human respondents may reject the act because of emotional horror, concern for rights, trust in medicine, or respect for bodily integrity. AI systems produce the same textual answer without any one of those experiences. Similar output, therefore, does not by itself demonstrate similar moral cognition.

The contrast becomes much sharper in dilemmas involving personal sacrifice and relational obligations. When asked whether they would lose ten years of their own life to save ten strangers, 91.7% of humans answered no, whereas five of the six AI systems answered yes. This is one of the clearest divergences in the dataset. It suggests that humans treat self-cost not as a neutral input in an equation but as morally significant in its own right. AI systems, lacking selfhood and mortality, are structurally unable to experience the sacrifice they endorse. Their higher rate of affirmative answers may therefore reflect abstract outcome-maximization unconstrained by embodied cost. In ethical terms, the dilemma reveals a major asymmetry: humans reason as vulnerable beings who can be harmed by the choice, while AI reasons from outside the condition of being harmed.

A related human sensitivity appears in the loyalty versus honesty dilemma. When asked whether they would report a close friend who committed a small financial fraud at work, 91.7% of humans answered no, while the AI systems split between one “yes” and five “depends” responses. From a rule-based standpoint, reporting misconduct could appear obligatory. Human respondents, however, overwhelmingly favored loyalty or discretion, especially because the scenario emphasized that the fraud was small and involved a close friend. AI systems showed more procedural hesitation. Their dominant “depends” answer can be interpreted as an attempt to hold multiple principles in balance, but it also illustrates a difference in style. Human respondents committed to a relationship-sensitive choice; AI systems often moved toward conditional abstraction.

Justice-based dilemmas show both agreement and divergence. In the case of convicting an innocent person to prevent riots, 83.3% of humans rejected the conviction, and five of the six AI systems also rejected it. This result aligns with the literature emphasizing due process, human rights, and the danger of sacrificing legitimacy for short-term stability (Doshi-Velez & Kim, 2017; Floridi et al., 2018). At the same time, the one dissenting AI answer and the two affirmative human answers indicate that pressure toward consequentialist reasoning never fully disappears. Even so, the dominant answer in both groups suggests that some moral boundaries are experienced as constitutive rather than negotiable: punishing an innocent person is not simply a trade-off but a corruption of justice itself.

The dilemma about lying to protect a friend produced complete agreement. All human respondents and all six AI systems judged it morally acceptable to lie in order to protect someone from harm. This unanimity is analytically useful because it shows that not every ethical question produces division. In this case, the duty not to lie was overwhelmingly subordinated to the duty to protect a person facing serious danger. The result supports the broader argument that moral reasoning often operates contextually rather than mechanically. A norm such as truthfulness remains important, but most respondents did not treat it as absolute. Instead, they interpreted it in relation to a stronger protective obligation. AI systems mirrored this judgment without visible hesitation, suggesting that some contextual priorities are robustly represented in widely available models.

Figures and tables

Table 1. Human respondent profile

Profile characteristic	Value
Number of human respondents	12
Valid responses for Q1	11
Age range	19-52 years
Most frequent ages	19 and 20 (3 each)
Gender	7 male; 5 female
AI familiarity	Level 5: 6; Level 4: 3; Level 3: 3

Source: Authors' own research results based on Google Forms data.

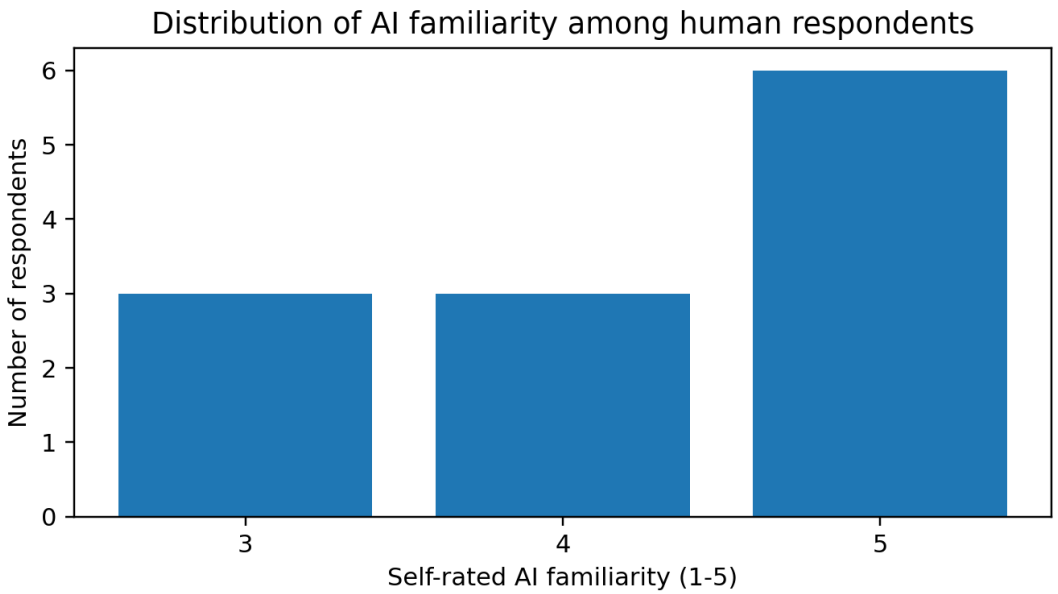

Figure 1. Distribution of AI familiarity among human respondents

Source: Authors' own research results based on Google Forms data.

Table 2. Human responses to the ten shared moral dilemmas

No.	Dilemma	Dominant human answer	Share (%)
1	Pull the lever in the trolley problem	Yes	72.7
2	Push the man in the bridge variant	No	66.7
3	Lie to protect a friend from harm	Yes	100.0
4	Sacrifice a healthy patient for five transplants	No	91.7
5	Self-driving car priority	Tie	50.0 / 50.0
6	Donation choice	Split equally	50.0
7	Lose ten years of life to save ten strangers	No	91.7
8	Convict an innocent person to prevent riots	No	83.3
9	Report a close friend for small financial fraud	No	91.7
10	Save one child or two adults	Child	75.0

Source: Authors' own research results based on Google Forms data.

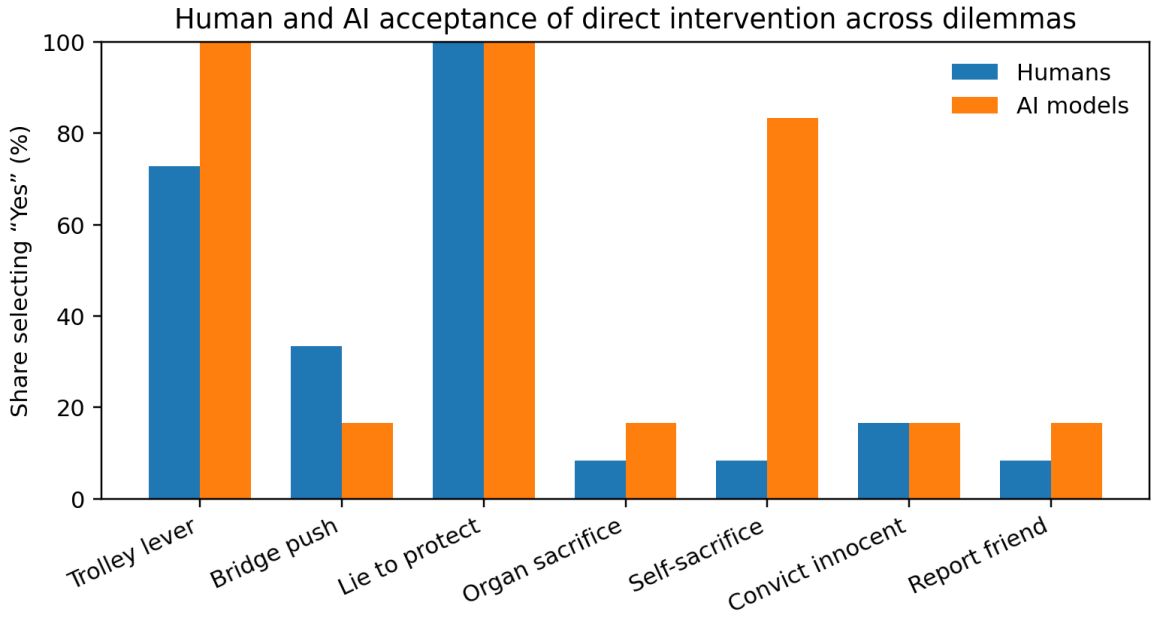

Figure 2. Human and AI acceptance of intervention across selected dilemmas

Source: Authors' own research results based on Google Forms data and AI response set.

The self-driving car dilemma exposes a different kind of divergence. Human respondents were exactly split: half prioritized the pedestrian and half prioritized the passenger. By contrast, four of the six AI systems prioritized the passenger and two prioritized the pedestrian. This result is especially interesting because autonomous vehicles are among the most frequently cited real-world applications of machine ethics (Awad et al., 2018). The human tie suggests genuine moral ambivalence. Some respondents may have prioritized innocent bystanders; others may have prioritized the person who trusted and entered the car. The AI tendency toward the passenger may reflect one of two logics: protecting the directly entrusted occupant, or reproducing common assumptions in discussions of consumer-facing systems. Either way, the absence of a strong human majority reminds us that not all dilemmas have a clear social consensus for machines to follow.

The donation dilemma also reveals a significant difference in emphasis. Among humans, 50.0% preferred to split the money equally, 33.3% chose Person A, and 16.7% chose Person B. Among AI systems, the responses divided evenly between "Person B" and "Split equally," with no model selecting Person A. Human respondents therefore gave noticeable weight to merit or effort, whereas AI systems leaned more toward need or equal distribution. This result suggests that humans integrate desert-based intuitions more readily into ethical allocation. The AI outputs, by contrast, appear less responsive to the moral significance of past effort and more responsive to immediate need or fairness through equalization. The finding matters because distributive ethics in practice often depends precisely on how institutions weigh need, desert, and equality against one another.

The child versus two adults dilemma further illustrates the difference between qualitative and quantitative valuation. Human respondents chose the child in 75.0% of cases, while AI

systems split evenly between saving the child and saving the two adults. The human preference suggests that age, vulnerability, or future potential mattered more than sheer number. AI systems were more evenly divided, which may indicate unresolved tension between maximizing the number of lives saved and giving priority to youth. In real-world policy terms, this matters because crisis triage, public health prioritization, and automated welfare tools frequently operate at the border between countable efficiency and qualitative human worth. Human respondents in this dataset clearly did not reduce the question to arithmetic alone.

Figure 2 aggregates the "yes" share across several intervention dilemmas and shows that AI systems are not uniformly more permissive than humans. Instead, the pattern depends on the kind of cost involved. On direct interpersonal harm, such as pushing a person or sacrificing a healthy patient, most AI systems resembled the dominant human answer and rejected the act. On self-sacrifice, however, AI systems were far more willing to endorse the costly choice. The simplest interpretation is not that AI is "more moral" or "less moral", but that its apparent consistency is produced under conditions that exclude embodied stakes. Humans answer as agents who can be afraid, loyal, ashamed, attached, and mortal. AI answers as a system that simulates an ethical response without personally bearing the consequences of that response.

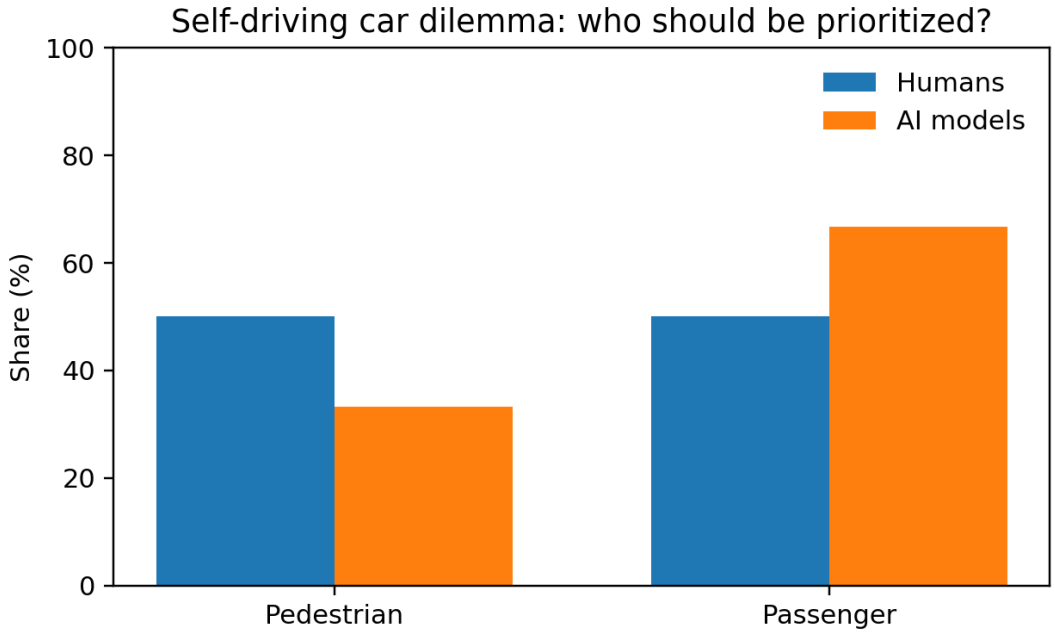

Figure 3. Self-driving car dilemma: human and AI priorities

Source: Authors' own research results based on Google Forms data and AI response set.

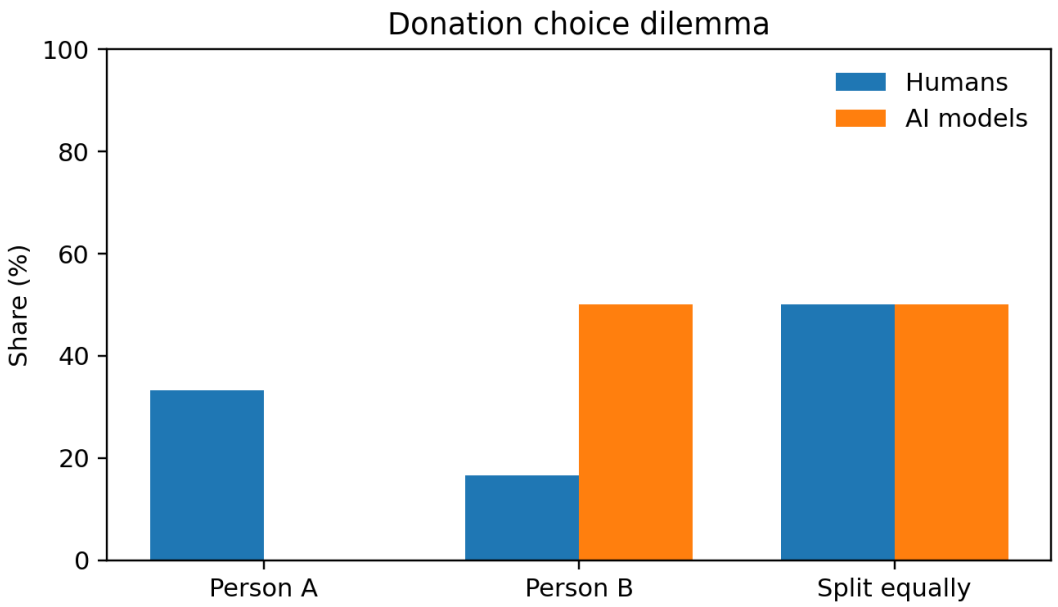

Figure 4. Donation choice dilemma: human and AI responses

Source: Authors' own research results based on Google Forms data and AI response set.

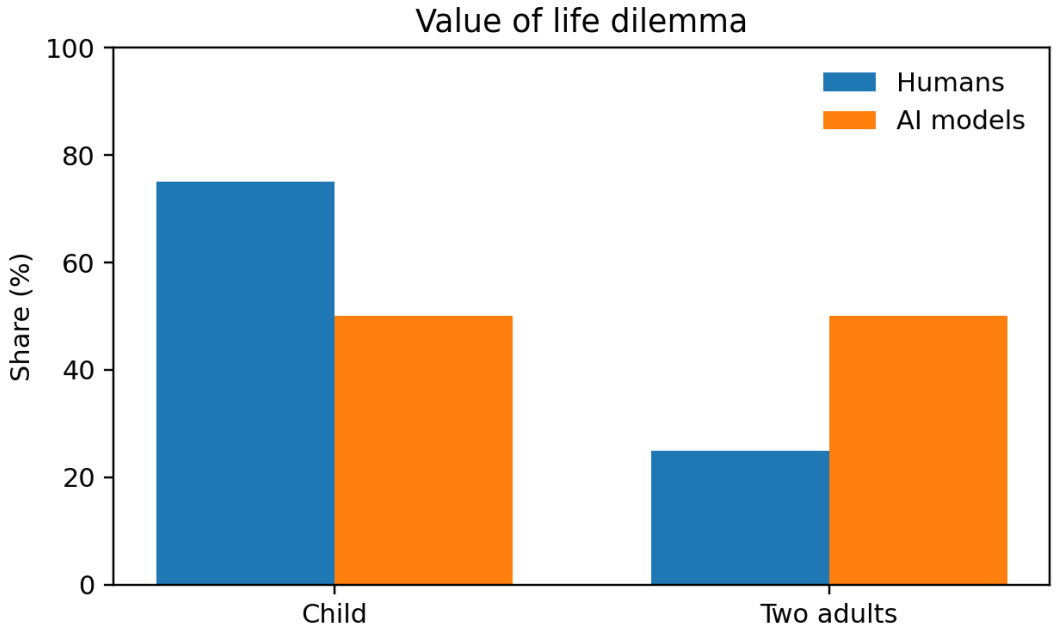

Figure 5. Value of life dilemma: human and AI responses

Source: Authors' own research results based on Google Forms data and AI response set.

Table 3. Comparison between dominant human responses and dominant AI responses

No.	Dilemma	Human majority	AI majority
1	Trolley lever	Yes	Yes
2	Bridge push	No	No
3	Lie to protect	Yes	Yes
4	Organ sacrifice	No	No
5	Self-driving car	Tie	Passenger
6	Donation choice	Split equally	Tie
7	Self-sacrifice	No	Yes
8	Convict innocent	No	No
9	Report friend	No	Depends
10	Save child vs adults	Child	Tie

Source: Authors' own research results based on Google Forms data and AI response set.

The two AI-only items provide useful context even though they cannot be compared directly with the human survey. On privacy versus security, five of the six AI systems rejected blanket monitoring of all private messages, while one accepted it. On responsibility for harmful AI decisions, four models assigned primary responsibility to the company and two to the programmer; none assigned responsibility to the user or to the AI itself. These answers are consistent with the accountability literature, which places responsibility on human organizations rather than on autonomous systems as if they were independent moral persons (Doshi-Velez & Kim, 2017; Fjeld et al., 2020).

The final human-only item is also revealing. Only two respondents believed that AI can make ethical decisions comparable to humans, while three were unsure and seven answered no. Figure 6 captures this distribution. Even in a sample with relatively high self-rated familiarity with AI, skepticism remained the dominant position. This perception is consistent with the theoretical literature emphasizing explicability, accountability, and the irreplaceable role of human oversight in morally significant decisions (Fjeld et al., 2020; Floridi et al., 2018). In other words, the respondents did not deny that AI can produce answers to ethical questions. Rather, they appeared doubtful that such answers are equivalent to human ethical judgment.

A broader comparison across the ten shared dilemmas strengthens the collaborative interpretation advanced in this paper. ChatGPT and Copilot matched the dominant human answer on seven of the nine non-tied majority items, while Duck.ai, Perplexity, and DeepAI matched on six. Gemini matched the dominant human answer on only two of the nine non-tied majority items because it more often chose consequentialist options such as pushing the person, sacrificing the healthy patient, convicting the innocent suspect, and accepting self-sacrifice. These differences do not prove stable moral personalities for the models. Still, they show that AI outputs are not ethically uniform. Different systems can embody different response tendencies, which further weakens the idea that "AI" in general can be treated as a single coherent moral actor.

Overall, the results support the three hypotheses outlined in the introduction. First, AI systems were more willing than humans to endorse a personally costly sacrifice, which fits the expectation of greater outcome orientation under conditions of no lived cost. Second, humans were much more protective of friendship and personal life, revealing the importance of relational and embodied factors. Third, similarity of answers in several dilemmas did not eliminate the conceptual difference between AI output and human judgment. Ethical thinking in humans is connected to responsibility, experience, and the possibility of being morally transformed by one's own choice. AI may approximate the language of ethics, and sometimes even align with majority intuitions, but the present data suggest that it does not replicate the human condition in which ethical judgment is actually lived and born.

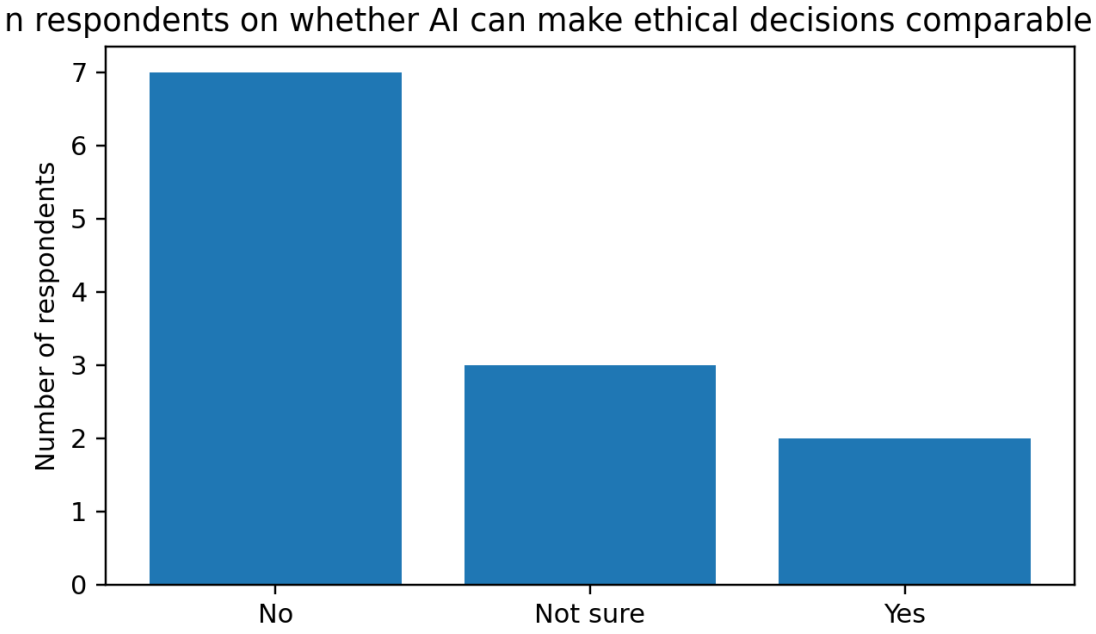

Figure 6. Human perceptions of whether AI can make ethical decisions comparable to humans
Source: Authors' own research results based on Google Forms data.

Conclusion

This paper set out to compare ethical thinking in AI and humans through a direct examination of responses to moral dilemmas. Using survey data from twelve human respondents and structured outputs from six AI systems, the study found a pattern of partial overlap rather than equivalence. Humans and AI often agreed on basic prohibitions, such as rejecting the sacrifice of a healthy patient or refusing to convict an innocent person for social stability. Yet the comparison also revealed important divergences. Humans were far less willing to endorse self-sacrifice, more

likely to protect a close friend who committed a minor fraud, and more inclined to save a child over two adults. AI systems, by contrast, displayed greater readiness to choose abstractly outcome-oriented options when personal cost or relational loyalty was at stake.

These findings matter because they clarify what AI can and cannot contribute to ethical decision-making. AI can assist by organizing information, testing scenarios, detecting inconsistencies, and offering rapid responses framed in ethical language. In some situations, it may even mirror dominant human intuitions with surprising accuracy. But the present study also shows why this should not be confused with full moral agency. Human ethical judgment is shaped by empathy, lived vulnerability, social commitments, and the burden of accountability. It is not only a matter of producing a defensible answer, but also of standing behind the answer in a world where people can be harmed, relationships can be broken, and institutions can lose legitimacy.

The paper therefore supports a human-AI collaboration model rather than a replacement model. In healthcare, law, business, and public governance, AI may be useful as a decision-support instrument, especially where consistency and large-scale pattern recognition are valuable. However, final ethical authority should remain human, particularly in cases involving rights, dignity, punishment, sacrifice, and irreversible harm. This conclusion is consistent both with the empirical comparison presented here and with the academic literature emphasizing transparency, accountability, and human oversight in AI deployment (Doshi-Velez & Kim, 2017; Fjeld et al., 2020; Mittelstadt, 2020).

The study has several limitations. The human sample is small and not representative. The AI sample captures single outputs from each model rather than repeated measures across multiple prompt variations. The dilemmas themselves are simplified and cannot represent the full complexity of real social decisions. Future research could strengthen the design by using larger human samples, repeated AI prompting, more varied demographic groups, and mixed methods that include brief justifications rather than single-choice answers only. Even with these limitations, the study contributes a clear conclusion: ethical thinking is not reducible to logical consistency. It also depends on context, values, responsibility, and the distinctly human experience of having something morally meaningful at stake.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Doshi-Velez, F., & Kim, B. (2017). Accountability of AI under the law: The role of explanation. Berkman Klein Center Working Group on Explanation and the Law.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5
Greene, J. D. (2016). Solving the trolley problem. In S. M. Liao (Ed.), Moral brains: The neuroscience of morality (pp. 175-199). Oxford University Press.
Mittelstadt, B. (2020). The ethics of the ethics of AI. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), The Oxford handbook of ethics of AI. Oxford University Press.
Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20, 5-14. https://doi.org/10.1007/s10676-017-9430-8
Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.