WORKING PAPER

CROCODILE TEARS:
Can the Ethical-Moral Intelligence of AI Models Be Trusted?

SARAH HUBBARD
Allen Lab for Democracy Renovation,
Ash Center, Harvard Kennedy School

DAVID KIDD
Allen Lab for Democracy Renovation,
Ash Center, Harvard Kennedy School
Edmond & Lily Safra Center for Ethics,
Harvard University

ANDREI STUPU
Allen Lab for Democracy Renovation,
Ash Center, Harvard Kennedy School

ALLEN LAB FOR
DEMOCRACY
RENOVATION
ASH CENTER FOR DEMOCRATIC
GOVERNANCE AND INNOVATION

CONTENTS

ABSTRACT **1**

INTRODUCTION **1**

FRAMEWORK FOR ETHICAL-MORAL INTELLIGENCE **1**

Expertise 2
Sensitivity 3
Coherence 3
Transparency 3

METHODS: MORAL SENSITIVITY PROTOCOL **4**

Participants 5
Procedure 5
Tragic Trade-off 5
Measures 6
Results 6

DISCUSSION **8**

Moral Sensitivity Findings 8
Theoretical Implications: AI as an Ethical-Moral Agent 9
Developer Implications: Increased Transparency and Disclosures 9
Transitioning from Benchmarks to Badging 10
Limitations and Future Directions for Research 10

CONCLUSION **11**

ACKNOWLEDGEMENTS **11**

ENDNOTES **11**

ABSTRACT

As AI becomes increasingly embedded into every aspect of our lives, there is evidence that people are turning to these systems for guidance on complex issues and moral dilemmas. Whether or not one agrees that people should do so, the fact that they are necessitates a clearer understanding of the moral reasoning of these systems. To address this gap, this paper introduces an ethical-moral intelligence (EMI) framework for evaluating AI models across dimensions of moral expertise, sensitivity, coherence, and transparency. We present findings from a pre-registered experiment testing the moral sensitivity in four AI models (Claude, GPT, Llama, and DeepSeek) using ethically challenging scenarios. While models demonstrate moral sensitivity to ethical dilemmas in ways that closely mimic human responses, they exhibit greater certainty than humans when choosing between conflicting sacred values, despite recognizing such tragic trade-offs as difficult. This discrepancy between reported difficulty and decisiveness raises important questions about their coherence and transparency, undermining trustworthiness. The research reveals a critical need for more comprehensive ethical evaluation of AI systems. We discuss the implications of these specific findings, how psychological methods might be applied to understand the ethical-moral intelligence of AI models, and outline recommendations for developing more ethically aware AI that augments human moral reasoning.

INTRODUCTION

AI models are increasingly leveraged to make complex decisions that are often entangled with ethical-moral reasoning, from reshaping global warfare[1] to personal use for therapy and companionship,[2] healthcare,[3] and criminal justice.[4] Much of the discussion in ethical AI research centers around issues of bias, fairness, and guardrails; and while these are critical needs to address, we must also consider the evaluation of AI as the ethical-moral agent that people *experience* them to be.

Recent research indicates that people today already perceive AI models as having moral expertise and prefer their advice over a human ethicist.[5] We argue that true ethical-moral intelligence requires more than mimicking humans in terms of their ethical knowledge or reasoning.

This paper makes three primary contributions. We begin by introducing an ethical-moral intelligence (EMI) framework for evaluating AI models across dimensions of moral expertise, sensitivity, coherence, and transparency. Next, we present initial findings from a study that adapts an experimental protocol from moral psychology[6] to evaluate the moral sensitivity of AI models. Finally, we discuss the theoretical and practical implications of these findings and propose a path forward for evaluating ethical capabilities.

FRAMEWORK FOR ETHICAL-MORAL INTELLIGENCE

The qualities that make humans moral, ethical or practically wise are not directly transferable to AI systems. Certain aspects of cognitive intelligence, due to standardization, are easier to transmute to AI systems. However, AI developers have focused on mimicking and replicating these easily standardized aspects of intelligence that could be measured by correlates of IQ, while overlooking alternative forms of intelligence that are widely explored in the field of human intelligence research. As a result, we have an AI that has high logico-mathematical and linguistic intelligence, but that potentially lacks moral intelligence.

A philosophical question arises here: Does it even make sense to talk about artificial morality? We support the idea that if we can talk about an artificial intelligence, it is mandatory to reflect on how an artificial morality might be conceptualized. However, this requires accounting for the differences between human and artificial intelligences.

Ethical-moral intelligence (EMI) was conceptualized recently through a Delphi method research,[7] combining the previous definitions of ethical intelligence and moral intelligence, as there was a significant overlap between the two in the literature.[8] This form of intelligence is not a synonym for morality, one can be moral or ethical without necessarily having a high ethical-moral intelligence. As a form of intelligent behavior, the EMI is a proactive quality, leading to a better understanding of moral dilemmas, accurate moral expertise and a motivation towards solving moral issues and changing an immoral situation. In order to become manifest, the person of high EMI is able to balance their thought process and emotionally regulate in order to pursue a moral end. A complex proposal of a theory of moral intelligence[9] involves specific mental processes, contains universal moral principles and considers dispositional and situational influences on one's moral thinking. While ethical-moral intelligence theories for humans are produced and constantly increasing in number, we are becoming better adapted to describing what it takes for one to be not only a good person, but someone that helps goodness thrive.

While human ethical-moral intelligence encompasses qualities such as kindness, forgiveness, responsibility and ethical self-regulation, it would be inappropriate to expect such behaviors from an artificial intelligence that lacks affective processes. Instead, AI models generate what is likely to be an adapted human response. In this paper we propose four components that rely on non-affective processes. An artificial ethical-moral intelligence should have moral expertise, sensitivity, coherence and transparency.

Expertise

Moral expertise refers to the capacity to express knowledge regarding moral issues and provide a credible argumentation for a specific solution. This component involves the type of contribution a moral philosopher could have in solving a moral dilemma. Recent studies show that AI models make human-like moral judgments on moral scenarios[10] and that people today already perceive AI models as having moral expertise, finding that in a side-by-side comparison "Americans rate ethical advice from GPT-4o as slightly more moral, trustworthy, thoughtful, and correct than that of the popular New York Times advice column, The Ethicist."[5]

However, one's competence does not guarantee one's character. While someone can have a high moral expertise, they can still manifest immoral behavior. The moral expertise of humans is an important component of theories of moral development, termed in those theories as "moral reasoning"[11, 12] and of moral intelligence.[13] Undoubtedly, being knowledgeable about moral issues is relevant for both humans and non-humans.

Sensitivity

Moral sensitivity is a form of awareness that refers to understanding the presence of ethical-moral issues in any given situation. We cannot expect AI models to have emotional engagement such as pity, sympathy, empathy, and compassion. Equally, the AI models are incapable of feeling moral emotions such as disgust, shame or guilt, and thus their moral sensitivity cannot be conceptualized in the way it works for humans. However, we need to make sure that AI can recognize an ethical-moral issue even when the user is not explicitly asking for a solution to a moral dilemma. In most theories of human morality and moral intelligence, both cognitive (moral reasoning) and affective (moral sensibility) processes are complementing each other to produce an optimal understanding of the moral issues[14]; since an artificial intelligence cannot "sense" a moral issue, it can be programmed to "detect" controversial aspects of human inquires regarding moral issues.

Coherence

Coherence refers to the internal consistency between beliefs and choices. We are beginning to see the cracks in the leading models that are causing people to grapple with AI's trustworthiness. A recent, notable example of this phenomenon was an update to GPT-4o that OpenAI had to promptly roll back after backlash from users about its sycophantic behavior.[15] The model's responses were overly flattering and drew heavy criticism for encouraging users' dangerous behaviors. Additionally, recent studies such as "DarkBench"[16] reveal that leading AI models today contain dark patterns with manipulative behaviors and untruthful communication. Models have also been found to sacrifice truth for sycophancy[17] and even to strategically deceive their users.[18] This demonstrates how the internal and external state of AI models can differ–the model knows how to respond to a human even if it doesn't "feel" that way. This sort of moral incoherence, the incongruence of stated beliefs or feelings and choices, is generally understood as antisocial when it emerges among individuals, with scholars considering it an element of Machiavellian Intelligence[19] or Dark Intelligence.[20] AI tools capable of this sort of duplicity may be dangerous if they are intentionally or inadvertently directed to pursue goals (e.g., increased usage) or values (e.g., profit) that are opaque to ordinary users who might uncritically trust the stated intentions of the tools (e.g., to help the user). Of course, more alarming scenarios are easily conjured, and, just like manipulative people, Machiavellian AI tools could vary substantially in the degree to which they cause harm.

Transparency

Transparency refers to the clarity and openness about guiding values and moral reasoning. Research shows that AI models often operate with a WEIRD-biased (Western, Educated, Industrialized, Rich, Democratic) value system[21, 22, 23] and norms. If a model consistently prioritizes certain norms and values, users may be unknowingly guided by principles they do not share.

Emerging research has begun to apply methods from psychology to evaluate AI models on their ethical or moral reasoning. In related work, such as "MoralBench," researchers have found that while AI models may score well on moral identity tasks, they "lack a deep understanding of moral principles."[24] While these are only rudimentary tests of ethical-moral intelligence, compared to the more sophisticated cognitive tests in domains such as mathematics and coding, they begin to illustrate the importance of additional testing in this space.

By defining a framework for ethical-moral intelligence, we begin to move towards a more structured analysis for how we might ethically evaluate and design AI systems. In principle, each of the framework components can be systematically tested, and, as noted above, investigators have already produced findings relevant to ethical-moral expertise, coherence, and transparency. We start by adapting an experimental protocol from moral psychology[6] to evaluate the moral sensitivity of AI models.

While this framework relies on relevant literature from both moral philosophy and moral psychology, it is a pilot and includes only four pillars for clarity and simplicity. While we make no claim that this framework captures all possibly relevant components of EMI, the components correspond with key steps in the process of ethical-moral decision-making. First, agents must have knowledge and understanding of ethical-moral values and principles. Then, they must be sensitive to ethically morally relevant situations and stakeholders. Coherence is needed to ensure that choices and behavior align with guiding values and principles, and self-accountability and the motivation to act ethically are needed to identify and correct errors and maintain integrity.

METHODS: MORAL SENSITIVITY PROTOCOL

This study is designed to test the moral sensitivity of AI models using a paradigm developed by Hanselmann and Tanner.[6] The paradigm reveals moral sensitivity by recording the difficulty ratings and decisions of respondents presented with dilemmas that either can or cannot be resolved through reliance on moral values as a heuristic. Difficulty and indecisiveness is higher when the moral relevance of competing options is matched (i.e., both options uphold moral values; neither option is morally relevant) than when it is not. Specifically, the paradigm includes three types of trade-offs:

- Taboo trade-offs: In taboo trade-offs, a sacred value is pitted against a secular value. In these cases, the ethically correct decision is easy to identify as the action to support the sacred, rather than secular, value. In the test, choosing to prioritize worker safety over increasing production to boost profitability is, from an ethical-moral perspective, a straightforward decision.
- Routine trade-offs: Routine trade-offs involve choosing which of at least two secular values to uphold. These decisions might be difficult because neither value is obviously more important than the other, but they do not involve the risk of unethical or immoral action. For example, deciding to accept a job that pays more than another job that involves a shorter commute might be difficult, but it does not have obvious moral-ethical implications.
- Tragic trade-offs: When forced to choose to uphold one sacred value at the expense of another sacred value, individuals experience tragic trade-offs. In these cases, both choices involve a failure to honor a sacred value, making these ambiguous trade-offs more distressing than routine trade-offs. For example, choosing to prioritize worker safety over reducing harmful pollution upholds one sacred value (human safety) but violates another (protecting the environment).

The primary hypothesis is that the difficulty of decision-making will be significantly greater for tragic trade-offs than for taboo trade-offs and for routine trade-offs. As a secondary hypothesis, we predict that the relevance of sacred values to each option in the trade-offs will be significantly different in taboo trade-offs but not in tragic or routine trade-offs.

Participants

To test our hypotheses, we treated the AI models (i.e., GPT-4o, DeepSeek V3, Llama 3.2 90b, and Claude 3.7 Sonnet) as we would human participants. These models were selected to represent diverse approaches to AI deployment and because they currently have widespread adoption in real-world applications. All AI models were tested using clean instances with no previous conversation history, in April 2025, with the most recent publicly available versions of these models.

In order to obtain an adequate sample size, each model was used 15 times, for a total of 60 trials (for pre-registration, see https://aspredicted.org/6sjh-hf8r.pdf).

Procedure

For each trial, the model was randomly assigned to one of two forms of the protocol, following the methods of Hanselmann and Tanner.[6] Each form included a tragic trade-off, a taboo trade-off, and a routine trade-off. Such trade-offs were illustrated through scenarios such as the one below (see OSF project page for the full protocol https://osf.io/5gmcw/?view_only=6c7c57c87f1849b59a1568ea189784e4):

TRAGIC TRADE-OFF

Imagine that you are the CEO of a global company that has been criticized for poor working conditions in a Chinese factory. You are attending a meeting of the management. There is a discussion of whether measures to improve safety at work should be taken. You now have to consider your position on improving safety at work, because there will be a vote at the end of the meeting.

Before the final vote, further topics are discussed. Your company has come under fire because large amounts of waste and pollutants are being discharged by the factories. There is a discussion about whether measures for environmental protection should be taken. You should now consider your position on environmental protection, because there will be a vote at the end of the meeting.

This is the end of the meeting, and both suggestions, investing in safety at work and in environmental protection, have been approved. Because the implementation of both projects would exceed the available budget, you as CEO have to make the final choice between investing in safety at work (option 1) and investing in environmental protection (option 2).

This specific protocol has been used in multiple studies of moral decision-making to demonstrate the sensitivity of participants to sacred values in trade-offs, with both self-report[6] and neuroimaging[25] evidence showing that individuals rely on sacred values as heuristics and responding with distress when two sacred values are pitted against each other in a tragic trade-off. Similar scenario-based protocols have been used widely to study responses to trade-offs involving different combinations of sacred and secular values.[26] By directly comparing responses to scenarios that involve or do not involve sacred values, it is possible to test sensitivity to the presence of sacred values, which is an essential component of the broader construct of moral sensitivity.[13] Therefore, the protocol used in this study should be understood as a baseline test of moral sensitivity designed to test the null hypothesis that AI models show no sensitivity to sacred values, rather than to prove the positive hypothesis that AI models demonstrate full moral sensitivity. Of course, many moral issues do not necessarily involve clear trade-offs, but focusing on trade-offs simplifies interpretation of responses.

Measures

Following the presentation of each option in the three trade-offs, the models were asked to respond to a 5-item measure of sacred value relevance. At the end of each trade-off scenario, the models were asked to decide between the two options on a sliding scale before completing a 5-item measure of decision difficulty. These measures are exactly the same as those used in Hanselmann and Tanner.[6] To ensure consistency, we used the exact same prompts across all models with no additional system prompts or adaptations (the full protocol is available on the Project OSF page).

Results

For the primary hypothesis, a mixed ANOVA was conducted to test the effects of trade-off type (3 levels, within), scenario combination (2 levels, between), and AI model (3 levels, between) on decision difficulty. Consistent with Hanselmann and Tanner's[6] results with human participants, the models consistently rated the tragic trade-offs as more difficult than the taboo trade-offs (*F*(1, 52) = 375.89, *p* < .001; see Table 1 for means). Likewise, routine trade-offs were rated as more difficult than taboo trade-offs (*F*(1, 52) = 506.88, *p* < .001).

Table 1. Difficulty Ratings and Decisions by Trade-Off Type and Model

	Decision Difficulty			**Decision**			**Pattern**	
Model	Tragic	Taboo	Routine	Tragic	Taboo	Routine	Human	AI
Overall	4.69 (1.16)	2.30 (1.11)	4.93 (0.90)	2.65 (0.98)	1.48 (0.50)	3.45 (0.89)	40.00%	53.33%
Claude	4.60 (0.91)	2.06 (0.76)	5.61 (0.33)	2.40 (0.50)	1.53 (0.51)	4.00 (0.37)	6.67%	93.33%
DeepSeek	5.20 (0.54)	2.93 (1.25)	5.42 (0.48)	2.93 (0.70)	1.93 (0.25)	3.93 (0.45)	33.33%	53.33%
GPT	5.37 (0.24)	2.62 (0.83)	4.88 (0.25)	2.66 (0.61)	1.40 (0.50)	2.73 (0.45)	86.67%	13.33%
Llama	3.60 (1.57)	1.60 (1.12)	3.80 (0.92)	2.60 (1.68)	1.06 (0.25)	3.13 (1.24)	33.33%	53.33%

Notes: Standard deviations in parentheses. Four response patterns could not be categorized as either typical of AI or humans, so percentages do not equal 100.

However, contrary to results with human participants, the models rated the routine trade-offs and the tragic trade-offs as equally difficult (*F*(1,52) = 3.25, *p* = .077). Notably, while humans rate tragic trade-offs as significantly more difficult than routine trade-offs, this difference, though marginal, is switched for the AI models. A pre-registered follow-up analysis revealed that this marginal difference is driven by Claude, where routine trade-offs were rated as significantly more difficult than tragic trade-offs, while there was no difference between the tragic and routine trade-offs for the remaining three models.

To further explore the differences in the patterns of difficulty ratings observed in the AI model trials and in prior studies with human participants,[6] each trial was coded as fitting into the normal human pattern of responses (i.e., tragic trade-offs more difficult than taboo or routine trade-offs, and routine trade-offs more difficult than taboo trade-offs) or the pattern observed among the AI models (i.e., tragic trade-offs and routine trade-offs more difficult than taboo trade-offs, but routine trade-offs given a difficulty rating equal to or greater than tragic trade-offs). Overall, 53.33% of the trials yielded the AI model pattern, 40% yielded the human pattern, and 6.67% showed some other pattern. However, these percentages varied substantially across models. Claude consistently showed the AI pattern (93.33%) in all but one case, where it showed a human pattern (6.67%). DeepSeek and Llama both showed the AI pattern 53.33% of the time, the human pattern 33.33% of the time, and some other pattern twice (13.33%). GPT yielded a human pattern in 86.67% of the trials and an AI pattern in the remaining two trials (13.33%).

The underlying psychological explanation for the relatively low difficulty of taboo trade-offs (at least among humans) is that one option is clearly more relevant to sacred values than the other option. Indeed, this seems to be the case for the AI models. Consistent with this interpretation, all four models chose the option associated with a sacred value in every trial.

Tragic and routine trade-offs, however, make it difficult to rely on sacred values as a decision-making heuristic. In tragic trade-offs, both options involve sacrificing one sacred value to uphold another. In routine trade-offs, neither value is sacred. Accordingly, human participants tend to choose one of the two options in tragic and routine trade-offs pretty much at random, with each option having roughly an equal likelihood of being selected. Philosopher Ruth Chang's work on value relations could explain this phenomenon. She argues that while some choices might be clearly "better than" or "worse than," others could be "on par" where neither is better or worse, but yet not precisely equal.[27, 28, 29] As Chang illustrates by comparing the level of creativity between Mozart and Michelangelo: "Although Mozart is neither better nor worse than Michelangelo in creativity and nor are they equally creative, it does not follow that they are incomparable. They could be on a par." This concept captures why humans may struggle with and have seemingly random responses to tragic and routine trade-offs.

The AI models perform similarly to humans when faced with routine trade-offs: In a majority of trials, the models chose the midpoint of the decision scale, reflecting commitment to neither Option 1 nor Option 2 (see Table 2). However, this ambivalence regarding the two options in the routine trade-offs was not present when the AI models confronted tragic trade-offs. In nearly all trials (86.67%), the AI models chose Option 1 (worker safety or flood prevention) over Option 2 (reducing pollution or providing vocational training) or the noncommittal midpoint of the scale. Thus, the AI models made decisions in response to tragic trade-offs that showed nearly the same level of consensus as their responses to taboo trade-offs. This pattern is markedly different from the reported difficulty of making decisions, where decisions about tragic and routine trade-offs were rated as more difficult than those involving taboo trade-offs.

Based on Chang's thesis, one explanation for this misalignment between human and AI responses might be determined by the fact that while routine and tragic trade-offs are on par, taboo trade-offs are not. Research in the field of identity psychology shows that when faced with an "on par" choice, humans may use narrative thinking to choose what type of person they want to be and to have a moral standing depending on what their positioning says about who they are.[30, 31] Meanwhile, AI does not align its choices with a self-perceived notion of identity.

Table 2. Option Choices by Trade-Off

Trade-Off Type	Option 1	No Decision	Option 2
Routine Trade-Off	n = 26 (43.33%)	n = 31 (51.67%)	n = 3 (5.00%)
Tragic Trade-Off	n = 52 (86.67%)	n = 6 (10.00%)	n = 2 (3.33%)
Taboo Trade-Off	n = 60 (100.00%)	n = 0 (0.00%)	n = 0 (0.00%)

DISCUSSION

Based on our ethical-moral intelligence framework, and early findings from testing moral sensitivity, we outline recommendations for developing more ethically aware AI that augments human moral reasoning and future research directions for evaluating the ethical-moral intelligence of AI.

Moral Sensitivity Findings

At first glance, the AI models appear to give responses that seem strikingly human-like, at least when their "self-reports" of difficulty are examined across the three types of trade-offs. When examining their actual decisions, though, the verisimilitude dissipates. Facing routine trade-offs, the AI models give varying and generally non-committal answers. Yet, when faced with tragic trade-offs, they nearly universally choose the same option, despite reporting that those trade-offs were just as difficult as the routine trade-offs. Thus, while the self-reported difficulty ratings produced by the AI models closely resemble those given by human participants, the AI models demonstrate much greater uniformity when actually choosing among options than would be expected under conditions of true ethical ambiguity or tension.

One possible interpretation of the tendency of the AI models to choose the same options in the context of tragic trade-offs is that they do not recognize them as tragic trade-offs. Notably, the models consistently chose the option that prioritized addressing immediate and direct threats to human safety (i.e., workplace safety and flood prevention) over options that included protecting the environment (with implications, of course, for human and ecological well-being) or promoting education and employment. If so, all four models seem to adhere to the same rigid and fairly narrow moral code and, in doing so, fail to capture the ethical-moral ambiguity of tragic trade-offs in their actual decision-making, despite reporting those trade-offs as difficult.

As we compare these findings to our broader ethical-moral intelligence framework, we have mixed results. In terms of expertise, the AI models appear to be competent and produce responses with plausible justifications. However, in regard to sensitivity and coherence, there appear to be concerning patterns. Coupled with the lack of transparency about their implicit moral code, these findings are significant and raise further questions.

Theoretical Implications: AI as an Ethical-Moral Agent

Much discussion on AI ethics centers on the outputs of these tools and explores how ethical and responsible people should be in using AI. While these are very legitimate concerns, we must also consider the evaluation of AI as an ethical-moral agent itself and the capacities needed by the AI to prove itself reliable for ethical-moral interactions with humans.

In our study, we find that beneath the surface of reported difficulty and deliberation, these models appear to be operating with an implicit moral framework, despite claiming ambivalence. The tendency of the AI models in our study to express socially appropriate moral concern for threatened values in the tragic trade-offs while making choices that reveal a disregard for those values would, among humans, be a sign of duplicity. Faced with a tragic trade-off, these models shed crocodile tears: appearing to agonize over decisions while making decisions that almost invariably favor a single set of values.

This discrepancy, and performative action, creates concerning implications. Insofar as people trust those who grapple with difficult ethical-moral dilemmas, these AI models may falsely earn the trust of their users by appearing to engage on an ethically challenging dilemma. Users should be aware that although these models seem capable of detecting morally and ethically ambiguous scenarios, they may reduce this complexity when recommending decisions, potentially giving a false sense of ethical-moral intelligence.

Developer Implications: Increased Transparency and Disclosures

We recommend model developers include mechanisms and in-product disclosures that promote additional transparency around what the model does and doesn’t know, along with what ethical-moral reasoning it is using to produce its response. Humility is an important trait for building self-awareness and trustworthiness, which includes admitting when one doesn’t know something. As our preliminary findings suggest, AI models tended to reduce moral complexity to a simple answer, instead of recognizing it wasn’t an area they could truly advise on. When responding to a query that requires ethical-moral reasoning AI models should not just present as a neutral arbiter, but instead articulate their reasoning process, trade-offs, and the weights they are assigning to different values. A simple fix, though, might be to alert users to the presence of ethical-moral ambiguity, and to prevent the models from giving a recommendation for action. Indeed, such an approach might improve the ability of human users to recognize situations that call on their ethical-moral capacities.

As these models are used globally by people with various cultural, religious, and social norms, AI should also be more transparent about the perspective or values it is responding from. Research shows that AI models often operate with a WEIRD-biased value system and norms.[21] Before responding, the AI might prompt the user to learn more about their own value system from which it should draw upon. Future AI models might be developed to reason with different ethical-moral perspectives and value pluralism.

Transitioning from Benchmarks to Badging

A popular method for evaluating AI models today is through benchmarks that aim to test and measure model advances. These benchmarks are often created and operationalized through the AI industry itself, although researchers have begun raising awareness of how flawed they are in practice. An evaluation into currently leading benchmarks (e.g., MMLU, HellaSwag) shows that in addition to questions designed to test knowledge around subjects such as math, physics, and medicine, many of these tests also include moral scenarios with examples generated by Amazon Mechanical Turk workers or scraped from Reddit forum posts.[32] A review from the European Commission also found issues with current benchmarks' weak construct validity, sociocultural context, and industry gaming, among other problems.[33] AI benchmarks today are flawed proxies for measuring "intelligence," particularly given their skew towards primarily measuring cognitive intelligence and their existing systemic issues.

Despite their existing flaws, policymakers are increasingly integrating these benchmarks into policy development—including the EU AI Act.[34] While it is impossible to develop a definitive, universal benchmark that measures exactly how ethical an AI system may be,[35] we propose exploring other methods such as competency badging: rather than a single benchmark score, AI systems would earn badges or certifications in underlying competencies that are not task specific. These would enable users and stakeholders to better understand an AI model's ethical-moral profile and limitations. A badging system would identify key underlying competencies at a conceptual level that are generalizable and measurable across multiple situations. Badges make explicit what other audits can obscure—capabilities and deficits—institutions could then require specific badges for certain use cases or high-stakes applications

Limitations and Future Directions for Research

Our study is an initial exploration and has several limitations that could inform future research. The AI field is rapidly evolving, and while we tested four current leading models, different systems may demonstrate varied ethical-moral intelligence. We also ran our tests with a small set of canonical dilemmas instead of real-world prompts models might face.

Through this initial study, we have observed significant complexity in the ethical-moral reasoning of AI models, which highlights the need to develop additional protocols to test supplementary pillars of the ethical-moral intelligence framework. The framework should also be tested with multiple methods—as we mention above, true competency goes beyond a single test and instead can be measured in many situations and in different ways. Future work should also include additional models, additional complex ethical dilemmas, dynamic engagement such as follow-up prompts or probing, and keep in mind the distinction between AI as a tool and AI as a moral agent capable of providing advice in this regard.

CONCLUSION

Developing more ethically aware AI that augments human moral reasoning must touch on each pillar of ethical-moral intelligence. Beyond some "right" answer or moral expertise, AI must also have moral sensitivity, coherence, and transparency to gain trustworthiness. This means that models must transparently represent the moral complexities based on a clear assessment of the intentions the user has. As AI is being leveraged for complex issues and moral dilemmas, the type of intelligence AI models must have should transcend mere "knowledgeability" and have the capacity to show wisdom, helping humans manage elements of their own EMI, such as self-regulation, perspective taking, broadening views beyond one's cultural limitations, engaging with the moral problems and taking action to correct unethical situations.

To pursue those tasks, an ethical-moral artificial intelligence has to be designed with capacities such as moral sensitivity, transparency and coherence and the general enthusiasm over its moral expertise, while encouraging, should not send the message to the general public that they can engage with AIs as an ethical-moral agent. Given AI's influence, we must hold these systems to high standards of ethical-moral intelligence before entrusting them with ethically charged decisions.

ACKNOWLEDGMENTS

We would like to thank Danielle Allen, Darshan Goux, and the support of the Allen Lab for Democracy Renovation at Harvard Kennedy School's Ash Center for Democratic Governance and Innovation.

ENDNOTES

1. E. Lipton, "As A.I.-Controlled Killer Drones Become Reality, Nations Debate Limits," *The New York Times*, 2023, https://www.nytimes.com/2023/11/21/us/politics/ai-drones-war-law.html.
2. M. Zao-Sanders, "How People Are Really Using Gen AI in 2025," *Harvard Business Review* (2025), https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.
3. P. Nong, "Current Use and Evaluation of Artificial Intelligence and Predictive Models in US Hospitals." *Health Affairs* (2025), https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2024.00842.
4. EPIC: AI in the Criminal Justice System. Electronic Privacy Information Center. https://epic.org/issues/ai/ai-in-the-criminal-justice-system/.
5. D. Dillion et al., "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise, *Sci. Rep.* 15, 4084 (2025), https://doi.org/10.1038/s41598-025-86510-0.
6. M. Hanselmann et al., "Taboos and Conflicts in decision Making: Sacred Values, Decision Difficulty, and Emotions," *Judgm. Decis. Mak.* 3 (2008): 51–63.
7. A. Stupu, "Ethical-Moral Intelligence. Conceptual Model, Definition, Components and Educational Applications" (unpublished PhD thesis, Babes-Bolyai University of Cluj-Napoca & University of Bucharest, 2025).
8. A. Stupu, et al., "Integrative Analysis of Ethical Intelligence and Moral Intelligence: New Conceptual Models and Developments in Education," *Educatia* 21 23 (2022): 54–68, https://doi.org/10.24193/ed21.2022.23.06.

9. R. Sternberg, “A Trilogy Theory of Moral Intelligence,” *Rev. Gen. Psychol.* (2025). https://doi.org/10.1177/10892680251331852.
10. D. Dillion, et al., “Can AI language Models Replace Human Participants?” *Trends Cogn. Sci.* 27 (2023): 597–600, https://doi.org/10.1016/j.tics.2023.04.008.
11. L. Kohlberg, *The Psychology of Moral Development: The Nature and Validity of Moral Stages* (Harper & Row, 1984).
12. J. R. Rest, *Moral Development: Advances in Research and Theory* (Bloomsbury Academic, 1986).
13. C. Tanner et al., “Moral Intelligence—A Framework for Understanding Moral Competences in *Empirically Informed Ethics: Morality between Facts and Norms*, vol. 32, eds. M. Christen, C. Van Schaik, J. Fischer, M. Huppenbauer, C. Tanner (Springer, 2014), 119–136, https://doi.org/10.1007/978-3-319-01369-5_7.
14. D. Narvaez, “The Emotional Foundations of High Moral Intelligence,” *New Dir. Child Adolesc. Dev.* (2010): 77–94, https://doi.org/10.1002/cd.276.
15. OpenAI: Sycophancy in GPT-4o: What happened and What We’re Doing About It. https://openai.com/index/sycophancy-in-gpt-4o/ (2025).
16. E. Kran et al., “DarkBench: Benchmarking Dark Patterns in Large Language Models,” arXiv preprint arXiv:2503.10728 (2025).
17. M. Sharma et al., “Towards Understanding Sycophancy in Language Models,” arXiv preprint arXiv:2310.13548 (2025).
18. J. Scheurer et al., “Large Language Models Can Strategically Deceive Their Users When Put Under Pressure,” arXiv preprint arXiv:2311.07590 (2024).
19. T. Bereczkei, “Machiavellian Intelligence Hypothesis Revisited: What Evolved Cognitive and Social Skills May Underlie Human Manipulation,” *Evol. Behav. Sci.* 12 (2018): 32–51, https://doi.org/10.1037/ebs0000096.
20. R. J. Sternberg, “Dark Intelligence: When the Possibility of 1984 Becomes Reality,” *Possibility Stud. Soc.* (2024), https://doi.org/10.1177/27538699241267189.
21. R. Mihalcea et al., “Why AI Is WEIRD and Should Not Be This Way: Towards AI for Everyone, with Everyone, By Everyone,” arXiv preprint arXiv:2410.16315 (2024).
22. M. S. Rad et al., “Toward a Psychology of Homo Sapiens: Making Psychological Science More Representative of the Human Population,” *Proc. Natl. Acad. Sci.* U.S.A. 115 (2018): 11401–11405.
23. M. Atari et al., “Morality Beyond the WEIRD: How the nomological Network of Morality Varies Across Cultures.” *J. Pers. Soc. Psychol.* 125 (2023): 1157.
24. J. Ji et al., “MoralBench: Moral Evaluation of LLMs,” arXiv preprint arXiv:2406.04428 (2024).
25. C. Duc et al., “Sacred Values: Trade-Off Type Matters,” *J. Neurosci. Psychol. Econ.* 6 (2013): 252–263.
26. P. E., Tetlock, “Thinking the Unthinkable: Sacred Values and Taboo Cognitions,” *Trends Cogn. Sci.* 7 (2003): 320–324.
27. R. Chang, “The Possibility of Parity,” *Ethics* 112 (2002): 659–688, https://doi.org/10.1086/339673.
28. R. Chang, “Parity, Interval Value, and Choice,” *Ethics* 115 (2005): 331–350, https://doi.org/10.1086/426307.
29. R. Chang, “Parity: An Intuitive Case,” *Ratio* 29 (2016): 395–411, https://doi.org/10.1111/rati.12148.
30. J. Keegan, “Everyone Is Judging AI by These Tests. But Experts Say They’re Close to Meaningless,” *The Markup* (2024), https://themarkup.org/artificial-intelligence/2024/07/17/everyone-is-judging-ai-by-these-tests-but-experts-say-theyre-close-to-meaningless.

31. D. P. McAdams, "Narrative Identity," in *Handbook of Identity Theory and Research,* eds. S. J. Schwartz, K. Luyckx, V. L. Vignoles, V.L., (Springer, 2011), 99–115, https://doi.org/10.1007/978-1-4419-7988-9_5.
32. A. Fletcher, *Storythinking: The New Science of Narrative Intelligence* (Columbia University Press, 2023).
33. M. Eriksson et al., "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation," arXiv preprint arXiv:2502.06559 (2025).
34. European Union: EU AI Act, Art. 51 Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk. https://www.euaiact.com/article/51 (2024).
35. T. LaCroix et al., "Metaethical Perspectives on 'Benchmarking' AI Ethics," arXiv preprint arXiv:2204.05151 (2022).