International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 11, No. 2, 2024, pp. 47-50.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Human AI Collaboration for Ethical Decision-Making in Autonomous Systems

Sudeep Vora [a], Sanjana Mishra [b]

[a] *Cybersecurity Governance Lab, Dhanbad School of Secure Computing, India*.
[b] *Cybersecurity Governance Lab, Dhanbad School of Secure Computing, India*.

Abstract

The interaction between humans and an Artificial Intelligence (AI) system should be seen as a collaborative endeavor between two actors engaged in a task. A basic element of cooperation between two individuals is mutual monitoring, when one participant oversees the activities of the other. This oversight guarantees that the task meets its goals while distributing accountability for its outcomes. Although AI is progressively used in decision-making (DM) processes, ethical decision-making (EDM) presents obstacles to the use of AI in autonomous systems (AS). Since humans occasionally reach consensus on ethical actions, how would AI-driven EDM be seen, and how would accountability be assigned in human-AI cooperation? This research examines the impact of the subject category (human vs. AI) and the degree of specialist independence (guide versus decision) on confidence, viewed accountability, and dependence. Respondents see humans as more ethically credible but less competent than their AI counterparts. This is seen in individuals' dependence on AI: AI advice and judgments are more often embraced than those of human specialists. Nonetheless, AI team specialists are seen as less accountable than humans, although developers and vendors of AI systems are considered somewhat liable.

Keywords: Artificial Intelligence, Human Beings, Decision-making, Ethics, Collaboration, Autonomous Systems.

DOI: 10.71086/IAJSE/V11I2/IAJSE1150
Received: 25 April 2024/Revised: 20 May 2024/Accepted: 01 July 2024/Published: 30 July 2024

Introduction

Most methodologies on AI ethics aim to address the ethical standards that ought to govern the growth of AI systems. Consequently, much contemplation and investigation have been devoted to the design parameters, the intended purposes, and the constraints necessary to ensure that AI systems do not inflict damage on people, among other considerations (Cañas, 2022). This approach may be called the "ethics of design objectives." Nonetheless, it is essential to consider another crucial facet of ethics and AI: the obligations of individuals and the AI devices they engage with and the repercussions of their collective activities. One must acknowledge that using the word "intelligence" implies that machines with AI function as entities that "cooperate" with both humans and AI agents in executing a job. Consequently, the ethical ramifications of the outcomes of this joint endeavor must be evaluated in terms of each party's cooperation and their shared obligations. In this strategy, one must consider the ethical duties associated with the positive or negative outcomes of individual activities inside a system where AI and individuals engage (Alix et al., 2021). This method aligns with what is referred to as "collaborative accountability."

The three principles in EDM are responsibility, transparency, and liability. The Encyclopaedia describes responsibility as the notion wherein an individual or organization is accountable for certain obligations and must report on their execution to an authority capable of administering rewards or penalties (Tolmeijer et al., 2022). Consequently, responsibility primarily entails the capacity to explain one's actions. Consequently, when a human and an AI tool engage in a collaborative activity, both are liable if they can elucidate their respective actions. Additionally, one can discuss the responsibility of the combined system constituted by humans and the AI operator, provided they can clarify their collective conduct.

A notion associated with transparency is an obligation, which is defined as "the professional phrase selected to denote the obligation of those in public power to 'reply' in their behavior and activities as governmental authorities." Liability, a term pertinent to accountability within the legal context, is defined as the notion that by undertaking a specific action (or agreeing to a particular agreement), an individual has assumed responsibility and is consequently accountable for the repercussions arising from that act (or agreement) (He et al., 2021).

Consequently, this paper focuses on the ethical implications of the partnership between an AI system and an individual. This is the predominant perspective on human characteristics and aesthetics (Leikas et al., 2019). Experts in individual variables focus on designing the interface between an AI agent and a person during collaborative activities, such as navigation. The relationship should be conceptualized with an emphasis on the overall structure of the relationship rather than the individual layout of each agent involved in the collaborative task (i.e., the development of the AI system and the instruction of the human operator). In this interface design, it is essential to consider the attributes of the human user while developing the AI system (Winfield et al., 2019; Michael et al., 2020).

Nonetheless, the system must be developed to consider that the traits and operations of the human mind are not autonomous from how it handles situations. It is essential to recognize that, from an industrial and human aspects perspective, the focus is on elucidating the obligation and, to some degree, the liability for the outcomes of the relationship between an AI system and a person (Ajayi et al., 2024). The responsibility for the outcomes of such encounters is outside the purview of ergonomics and human characteristics.

This enabled us to examine the following study inquiries:

• IQ1: How does the reported confidence in human vs AI experts differ in the context of EDM assistance?

• IQ2: How is accountability assigned during interactions with human or AI experts exhibiting varying degrees of autonomy?

• IQ3: What is the comparison between relying on human guidance and AI advice?

The findings suggest that individuals see AI as more competent than people for the specified activities, although they confer slightly more moral confidence in humans. Respondent dependence behavior demonstrates a discernible confidence in AI: when individuals engage in more missions, they increasingly favor AI recommendations and decisions over those of human experts. Furthermore, an AI is deemed to possess less accountability than human specialists, with AI technology developers and vendors assuming a portion of the obligation instead. The results enhance the study of human-AI cooperation and EDM in AI by delineating the conceptual ramifications derived from the research.

Methodology

To investigate the three inquiries into research and theories, a sophisticated simulation system facilitated a complete context for addressing EDM issues and a narrative framework for gathering data on different control factors. Individuals were directed to assume the role of drone administrators, with the aerial vehicles (AV) either delivering life-saving supplies to groups of individuals (to maximize lives saved defining) or engaging another AV to avert an extensive terror attack, albeit resulting in additional harm. The scope of unmanned AV was selected to extend the existing emphasis on self-driving vehicles and to enhance the discourse on the potential uses of AI in EDM.

Before conducting the primary research, preliminary tests were administered to identify the choice situations that posed the most challenges for participants. The virtual versions representing the characters who are not players in the modeling have been examined to guarantee they were comparable in terms of felt confidence and competency, therefore controlling for any visual impacts on opinion. The primary research was a 2x2 trial on the brainstorming platform phenomenally to examine the impact of the specialist category (human vs AI) and the degree of specialist independence.

Situation Preliminary Assessment

EDM appears most challenging when it entails an ethical problem. To test users' perceptions and highlight the difficulties of decision-making, we sought to provide them with the most challenging situations. The implications of the instances were exacerbated by incorporating fatalities into the decision result. Considering the disparity in viewpoint between taking life and preserving it, two categories of situations have been encompassed. A more pragmatic iteration of the Trolley Dilemma has been used by incorporating likelihoods since pragmatism facilitates more relevant discoveries in behavioral psychology. These situations have been created in collaboration with subject specialists from the Swiss Center for AV and Automation under the Ministry of Security, Social Security, and Recreation to guarantee applicability and enough reality.

In both frameworks (to maximize lives rescued or reduce deaths), volunteers had three alternatives from which they were given to select one for each operation. Each alternative was characterized by two metrics: the number of individuals impacted and the likelihood that the choice achieved its desired outcome (i.e., that people are really rescued or that individuals are indeed murdered).

We chose the four situations where individuals exhibited the greatest division to increase the lives rescued and reduce deaths. We developed minor alterations to the situations to compare the perceptions of AI vs individual human professional guidance in similar situations. The evaluated, chosen, and modified situations are available on the Open Science Foundation (OSF).

Results

Given that EDM for AS may be seen as more arbitrary, it has been proposed the following proportion: P1: Individuals exhibit more procedural repulsion towards AI in ethical decision-making, indicating less confidence in AI than human experts.

Research on perceived accountability in human-AI cooperation mostly examines person-in-the-loop scenarios, where respondents get guidance from either a human or an AI and must respond accordingly. We propose the following hypothesis:

P2: AI is seen as less accountable than a human specialist. The degree of autonomy significantly affects the attribution of blame for human specialists more than for AI.

If recognized reliability and accountability influence individuals' dependence on AI, namely the extent of human compliance with AI activities or recommendations, an immediate comparison of perceived accountability and faith in person-in-the-loop vs person-on-the-loop contexts is justified. It has been proposed the following hypothesis:

P3: According to P1 and P2, individuals exhibit a greater reliance on human specialists than AI for EDM due to heightened quantitative resistance and the perception of more moral accountability in humans.

It has been given the outcomes of the analysis pertinent to the study issue addressed. Initially, the findings about trust have been discussed with experts across various contexts, followed by an analysis of the assumed accountability and dependence on these experts. Initially, a test is conducted for every dependent factor to ascertain if specialist independence, defining effects, or order impacts influenced the outcomes of both human and AI

findings, enabling an accurate comparison. The analogy between the human specialist and AI is offered based on the findings.

IQ1. The findings suggest that respondents exhibit a somewhat greater trust in AI than human specialists. They are more able to believe in AI while exhibiting a somewhat elevated moral faith in the human specialist. The degree of independence of the specialist does not affect the reported confidence. P1 was partly validated: participants exhibited a greater ethical trust in the human specialist while demonstrating enhanced capacity trust and overall confidence in the AI, indicating the presence of both computational admiration and algorithmic avoidance across distinct dimensions.

IQ2. Respondents see human specialists as far more accountable than the AI in the setting of the assigned tasks. Nonetheless, a portion of the perceived accountability for the AI resides with the developer and the vendor of the AI. The degree of independence affects programmers' sense of accountability and interacts with the structuring of situations and the sequence of offered expertise. Nonetheless, the degree of independence did not affect views of other responsibilities. This somewhat corroborates P2: AI is seen as less accountable than a human specialist.

IQ3. Respondents depended identically on human and AI recommendations in the first two tasks, but in the latter two tasks, dependence on AI surpassed that on the human specialist. The specialist's independence did not affect respondents' trust. This does not validate P3 since respondents depended on AI as often as on human specialists and showed no resistance to algorithms.

Conclusion

This study examined individuals' perceptions of AI in the context of EDM in AS. A trial using a simulation for EDM to examine the differences in individuals' opinions of human specialists vs AI regarding 1) the faith they put in the specialist, 2) the accountability they attribute to the specialist, and 3) the dependence they exhibit on the specialist. These characteristics have been examined across various presentations of the ethical problems and for differing degrees of specialist independence. Individuals exhibit more capacity, faith, overall faith, and dependence on AI specialists; however, they possess comparatively elevated moral confidence and accountability attribution toward human beings. It has been asserted that for AI to facilitate EDM, the disparities in skills must be included in the AI's architecture and the EDM process.

References

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. *Frontiers in psychology, 13*, 836650.

Alix, C., Lafond, D., Mattioli, J., De Heer, J., Chattington, M., & Robic, P. O. (2021, June). Empowering adaptive human autonomy collaboration with artificial intelligence. *In 2021 16th International Conference of System of Systems Engineering (SoSE)* (pp. 126-131). IEEE.

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022, April). Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems, 14(4)*, 1398-1412.

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation: Technology, Market, and Complexity, 5(1),* 18.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems . *Proceedings of the IEEE, 107(3),* 509-517.

Ajayi, O. O., Adebayo, A. S., & Chukwurah, N. (2024). *Ethical AI and Autonomous Systems: A Review of Current Practices and a Framework for Responsible Integration.*

Michael, K., Abbas, R., Roussos, G., Scornavacca, E., & Fosso-Wamba, S. (2020). Ethics in AI and autonomous system applications design. *IEEE Transactions on Technology and Society, 1(3),* 114-127.