Artificial Intelligence, Social Computing and Wearable Technologies, Vol. 113, 2023, 69–76
https://doi.org/10.54941/ahfe1004180

Leveraging Multi-User Dungeons for Ethical AI Decision Support Systems: A Novel Approach

Daniel Pittman[1], Kerstin Haring[2], and Chris GauthierDickey[2]

[1]Metropolitan State University of Denver, Denver, CO 80204, USA
[2]University of Denver, Denver, CO 80210, USA

ABSTRACT

This paper proposes the innovative use of Multi-User Dungeons (MUDs) as a testbed for exploring and refining Artificial Intelligence (AI) ethics in decision support systems. MUDs are interactive, text-based virtual environments and offer a unique platform for studying AI behavior in a controlled yet complex environment. Our approach involves a combination of machine learning and natural language processing techniques to implement AI as a decision support system, and designs scenarios that challenge players with ethical quandaries and dilemmas. The effectiveness and ethical decision-making of players, the AI, and both together as a team are evaluated through a mix of quantitative and qualitative methods. The approaches detailed in this research aim to contribute to the broader discourse on AI ethics, stimulate a discussion on how to provide empirical evidence of AI decision-making's impact on human behavior in MUDs, and informing the design of ethically responsible AI systems in other domains.

Keywords: Artificial intelligence ethics, Multi-user dungeons (MUDs), Human-AI interaction

INTRODUCTION

The rapid evolution of artificial intelligence (AI) has amplified the use of AI-driven decision support systems across diverse sectors, including the military, healthcare, finance, and more (Russell and Norvig, 2016). As AI systems become more integrated into our daily lives, it becomes imperative to address the ethical dimensions of AI behavior to ensure responsible and equitable decision-making. Although AI systems can improve people's lives, they also can be deeply biased and lack transparency (Buolamwini and Gebru, 2019), and the growing awareness and AI literacy of especially younger generations calls for a broad adoption of ethical AI design (Casal-Otero et al., 2023).

AI ethics, a burgeoning field, seeks to understand and address the ethical implications of AI systems, including issues such as fairness, accountability, transparency, and potential harm (Bostrom, 2014). As AI systems become more autonomous, their decisions can significantly impact individuals and society, making it crucial to develop methods and frameworks for ensuring ethical AI behavior. Researching the ethical challenges to AI decision systems, ideally before there are real-world consequences of the decisions made by AI

and the acceptance and adoption of such decisions by humans is therefore crucial.

In this paper, we propose the innovative use of Multi-User Dungeons (MUDs) as a testbed for exploring and refining AI ethics in decision support systems. MUDs are interactive, text-based virtual environments where multiple users can engage simultaneously, offering a unique platform for studying AI behavior in a controlled yet complex environment (Bartle, 2003). For example, MUDs can combine role-playing games where people assume the role of a character in a fictional setting, hack and slash which emphasizes combat, Player vs. Player simulating conflict, interactive fiction where text is used to control characters and influence the environment, and online chat as real-time communication between players and characters. By incorporating AI-controlled non-player characters (NPCs) into MUDs, we can simulate a variety of scenarios reflecting real-world and fictional situations, providing a rich context for studying the ethical implications of AI decision-making.

The use of MUDs for studying AI ethics offers several advantages, including the study of AI behavior in a dynamic, real-time environment, detailed analysis of AI behavior and user interactions, and insights into how AI systems interact with humans in a collaborative setting.

This paper delves into the methodologies we propose for exploring AI ethics within MUD platforms, the potential applications of this research, and the broader implications for AI ethics. We detail our approach to integrating AI into MUDs, describe the testing scenarios and ethical decision-making considerations, and outline our evaluation methods. We conclude with a discussion of the insights gained from this research and their implications for the design of ethically responsible AI systems in other domains.

Background

Artificial Intelligence (AI) has been increasingly integrated into decision support systems across various sectors, including healthcare, finance, and the military (Russell and Norvig, 2016). These systems, powered by advanced machine learning algorithms, are capable of processing vast amounts of data and making complex decisions at a speed and scale beyond human capabilities. However, the ethical implications of AI behavior in these systems have been a growing concern (Bostrom, 2014), including the trustworthiness results (Durán and Jongsma, 2021), the reliability and transparency of black box algorithms (Von Eschenbach, 2021), and the desirable actions based on the results (Bélisle-Pipon et al., 2022).

For example, AI-based decision support systems have shown promising results in preclinical evaluations, but few have yet demonstrated real benefit to patient care (Vasey et al., 2022). The DECIDE-AI reporting guideline described by Vasey et al. provides a framework for reporting early-stage clinical studies of AI-based decision support systems, emphasizing the importance of assessing an AI system's actual clinical performance, ensuring its safety, and evaluating the human factors surrounding its use.

In the context of type 1 diabetes management, AI-based decision support systems have been used to deliver personalized recommendations regarding insulin doses and daily behaviors (Tyler and Jacobs, 2020). In supply

chain risk management, decision support systems have leveraged AI techniques such as Petri nets, multi-agent systems, automated reasoning, and machine learning to analyze data and make decisions regarding potential risks (Baryannis et al., 2019).

While these examples highlight the versatility and potential of AI in decision support systems across different sectors, they also underscore a critical gap in our understanding of the long-term consequences of AI recommendations. Specifically, there is a need to explore how humans react when they disagree with AI decisions and how these disagreements impact their trust in and use of AI systems. This is particularly relevant as AI systems become more integrated into our daily lives and decision-making processes. Moreover, the need to effectively and to correctly evaluate trust in AI-assisted decision-making requires well thought out protocols (Vereschak et al., 2021).

Transitioning from these real-world applications to Multi-User Dungeons (MUDs), the interactive, text-based, and multi-user engagement components (Bartle, 2003) offer a virtual environment where the impact of AI decisions and recommendations can be researched in-depth without the real-world consequences. Traditionally used in gaming, players interact with each other and with Non-Player Characters (NPCs) controlled by the game's AI. The use of AI in MUDs has been largely unexplored in the literature, presenting an opportunity for novel research.

Implementing and Evaluating AI in Multi-User Dungeons (MUDs)

Our approach to integrating AI into MUDs and studying its implications for ethical decision-making involves several steps, which are outlined in the following subsections. This process is iterative and involves continuous refinement based on the feedback and data gathered during the testing and evaluation stages.

Why Use MUDs Over Traditional Video Games?

Multi-User Dungeons (MUDs), with their text-based and interactive nature, offer a unique platform for studying AI ethics that traditional video games may not provide. The primary advantage of using MUDs is their focus on text-based conversation and interaction, which allows for a more in-depth exploration of AI decision-making and ethics. In a MUD, the quality of the interaction is not influenced by the quality of graphics or other visual elements, which can sometimes distract from the core gameplay and decision-making processes in traditional video games. This focus on text and interaction allows us to closely examine the ethical implications of AI decisions and the players' responses to them. Furthermore, the text-based nature of MUDs allows for faster iteration and development, as there is no need to invest in creating complex 3D environments or graphics. This enables us to adapt and refine our AI models and scenarios based on player feedback and research findings, accelerating the pace of our research into AI ethics more quickly.

Text-based interactions hold untapped potential for Natural Language Processing (NLP) as language not only reflects human behavior and individual traits, but also provides context about the author and situation. This

latent information can be leveraged by text-based technologies, necessitating an assessment of ethical implications (Hovy, 2016).

AI Integration and Ethical Considerations

The integration of AI into MUDs involves not only technical implementation but also the incorporation of ethical considerations. This is a critical aspect of our approach as it ensures that the AI does not merely function within the MUD environment, but also adheres to ethical guidelines that govern its interactions with players. Drawing from the work of Bostrom and Yudkowsky (2014), the AI is designed to provide decision support and challenge players with ethical dilemmas. These dilemmas, embedded in the scenarios, require players to make decisions that have moral or ethical implications. The AI responds to these decisions in ways that reflect different ethical frameworks, providing an opportunity for players to reflect on the consequences of their actions. This approach allows us to observe how players respond to ethical challenges in real-time and how their decisions influence the unfolding of the game narrative.

Decision Support System Implementation and Ethical Scenarios

The AI serves as a decision support system (DSS), providing players with suggestions and guidance based on their current situation in the game. Instead of using more traditional machine learning approaches to creating the AI model that the players interact with, such as reinforcement learning, we propose a novel approach that leverages the capabilities of large language models (LLMs) to create two distinct AI personas: a "good" persona and a "bad" persona.

Each persona is designed to suggest radically different methods to achieve the same goal, thereby presenting players with a range of ethical choices. The "good" persona suggests actions that align with generally accepted ethical norms, while the "bad" persona proposes actions that may be ethically questionable. This approach allows us to present players with a diverse set of ethical dilemmas and observe their decision-making processes (Bartle, 2004).

A large language model (LLM) is a type of machine learning model that can perform a variety of Natural language processing (NLP) techniques, such as sentiment analysis and topic modelling, are used to interpret player inputs and generate responses (Liu, 2012). This enables the AI to detect the nuances of the responses by players as it presents them with options and react to those nuances in real time with the LLM.

In this approach, the AI is designed to respond to the ethical dynamics of the current group of players in a manner consistent with its current persona. It does this by observing the players' responses to the suggestions made by the two personas. This allows the AI to gauge the ethical compass of the group and align its behavior and suggestions accordingly. For example, if a player strongly objects to a proposed action on ethical grounds, the AI, in its "good" persona, would respect this objection and adjust its future suggestions to avoid similar ethical conflicts. Conversely, the "bad" persona might challenge the player's objection, creating further ethical tension in the game scenario. This approach not only respects the player's ethical stance but

also provides a dynamic and engaging environment where players can experience the consequences of their ethical decisions in a safe and controlled setting.

This approach provides a more nuanced and personalized exploration of AI ethics, enhancing the realism and engagement of the game scenarios. By using LLMs to create distinct AI personas and respond to player decisions in real time, we can investigate the ethical implications of AI decision-making in a dynamic, interactive environment.

Testing Scenarios and Ethical Decision-Making

A variety of game scenarios, including combat, exploration, puzzle-solving, and social interaction, are designed to test the AI's ability to understand and respond to player inputs, adapt to player actions, and provide useful advice or suggestions. These scenarios also test the ethical decision-making of the players and the AI, with the AI programmed to respond based on a variety of ethical frameworks. The scenarios are designed to be challenging and engaging, encouraging players to think critically about their actions and the potential consequences. By observing how players navigate these scenarios, we can gain valuable insights into how they make ethical decisions and how these decisions are influenced by the AI's suggestions and guidance.

Player Engagement and Ethical Realism in MUDs

In order to ensure that players respond to ethical dilemmas in our MUDs as they would in real life, creating scenarios that are engaging and realistic is crucial. This involves designing ethical dilemmas that are not only challenging but also relatable and meaningful to the players.

A key aspect of this is providing players with the freedom to challenge the ethical nature of the AI's direction. This is exemplified in a controversy surrounding the quest "Torture the Torturer" and "The Art of Persuasion" in the World of Warcraft expansion Wrath of the Lich King, where players were required to torture a character for information. Richard Bartle, co-creator of the first MUD, expressed his displeasure with this type of quest, stating that he expected there to be a way for players to refuse to participate in the torture but found no such option (Engadget, 2008).

This incident highlights the importance of giving players agency in ethical decision-making within games. In our MUD, we aim to provide players with the ability to question and challenge the ethical decisions suggested by the AI. This not only enhances player engagement but also encourages players to reflect on their own ethical values and decisions.

Moreover, to ensure that players do not become detached from the ethical dilemmas and quandaries presented in the MUD, we aim to design scenarios that are closely aligned with real-world situations. This is supported by Bartle's (2003) argument that players are more likely to engage with ethical dilemmas in games when they perceive them as being relevant to their own lives. Often, ethical dilemmas are researched in extremes, utilizing variations of the trolley problem (Bonnefon et al., 2021; Awad et al., 2021). However, these extreme scenarios of life and death are unlikely to be a regular decision outcome of an AI decision. More likely, there are ethical quandaries, and

while the decision and outcome can have tremendous consequences on individuals' lives, they are not as severe as life and deaths decisions. For example, an "offer you can't refuse" would be a competing job offer with a significant raise from a competitor while one is employed by a supportive company that, however, pays less (Pastin, 2018). Another example, a "my back yard" scenario, includes a development that is great for the broader community, but not for the people near the development (like a community center, safe drug use centers).

Likewise, science fiction scenarios could be used to explore ethics in a MUD through reasoning how a new technology like robots should or should not be built or employed. For instance, a scenario could be created where players are part of a governing body deciding on the ethical guidelines for the creation and use of sentient robots. Players could be presented with different perspectives, such as the potential benefits of such technology for society (e.g., increased productivity, performing tasks humans cannot do) and the potential risks (e.g., job displacement, mistreatment of sentient beings). Another scenario could involve the use of AI in healthcare, where players must weigh the benefits of AI's efficiency and accuracy against potential issues like privacy concerns and the loss of human touch in care. These scenarios not only provide a platform for players to engage with ethical issues in a meaningful way but also allow the AI to learn from the players' decision-making processes and adapt its behavior accordingly.

Evaluation of Effectiveness and Ethical Decision-Making

The effectiveness of the AI and DSS is evaluated through a combination of quantitative and qualitative methods. Quantitative evaluation involves analyzing game metrics such as player success rates, decision-making times, and game completion times. These metrics provide a clear measure of the AI's performance and its impact on the player's experience. It would also include the result of sentiment analysis and topic modeling of the resulting player surveys where we ask for feedback regarding the scenario they participated in. Qualitative evaluation involves collecting player feedback through surveys and interviews, with specific questions about the ethical dilemmas in the scenarios. The ethical decision-making of the players and the AI is also evaluated, with decisions analyzed based on established ethical frameworks (Bostrom & Yudkowsky, 2014). In addition, the decision-making efficiency of the AI is evaluated by comparing the suggestions made by the two personas and the players' responses to these suggestions. This comprehensive evaluation approach allows us to assess the effectiveness of the AI and DSS from multiple perspectives and gain a deeper understanding of their impact on player behavior and decision-making.

Participant Questions and Ethical Feedback

Participants are asked questions about their experience with the AI and DSS, including their perceived usefulness, ease of use, and satisfaction. They are also asked about their understanding of the AI's decisions and suggestions, and their level of trust in the AI. Specific questions about the ethical dilemmas in the scenarios provide insights into how players navigate these dilemmas and how they perceive the AI's ethical decision-making. This feedback is

invaluable in refining the AI and DSS, ensuring that they are not only effective but also perceived as helpful and trustworthy by the players. It also provides insights into how players perceive and respond to ethical dilemmas, which can inform the design of future scenarios and ethical frameworks.

Statistical Testing and Ethical Insights

Statistical tests are used to analyze the results and determine the relevance of the findings. This includes tests to compare player performance with and without the AI and DSS, and tests to compare the performance of different AI and DSS implementations. The tests include paired t-tests to compare player performance with and without the AI, and correlation analysis to examine the relationship between player performance and their feedback about the AI (Field, 2013). The results of these tests provide insights into the factors that contribute to ethical AI decision-making in MUDs. They also highlight the strengths and weaknesses of the AI and DSS, providing a clear direction for future refinements.

FUTURE WORK

There are several promising avenues for future research related to this topic. One such avenue is the exploration of long-term effects of AI decision-making on player behavior. This would involve conducting longitudinal studies to observe how repeated interactions with the AI and exposure to ethical dilemmas influence players' decision-making patterns and ethical perspectives over time. Another potential area of future work is the expansion of this research to other interactive environments, such as virtual reality or augmented reality platforms. These environments offer even more immersive experiences and could provide additional insights into human-AI interaction and ethical decision-making. Furthermore, the potential for using MUDs as a tool for AI ethics education could be explored. This could involve developing educational modules or games that use the AI and ethical dilemmas to teach players about AI ethics and ethical decision-making. These future research directions have the potential to further our understanding of AI ethics and contribute to the development of more ethically responsible AI systems.

CONCLUSION

The integration of AI into Multi-User Dungeons (MUDs) offers a unique platform to investigate the ethical implications of AI decision-making in a dynamic, interactive environment. Our approach combines machine learning and natural language processing techniques to implement AI as a decision support system, and designs scenarios that challenge players with ethical dilemmas. The effectiveness and ethical decision-making of both players and the AI are evaluated through a mix of quantitative and qualitative methods. The findings from this research will contribute to the broader discourse on AI ethics, providing empirical evidence of AI decision-making's impact on human behavior in MUDs. This research has demonstrated the potential of MUDs as a tool for studying AI ethics, and the insights gained have significant implications for the design of ethically responsible AI systems in other domains.

REFERENCES

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2021). The Moral Machine experiment. Nature, 563(7729), 59–64.

Bélisle-Pipon, J. C., Monteferrante, E., Roy, M. C. and Couture, V., 2022. Artificial intelligence ethics has a black box problem. AI & SOCIETY, pp. 1–16.

Bartle, R. (2003). Designing Virtual Worlds. New Riders.

Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Decision Support for Risk Management in Global Supply Chains: A Multi-objective Approach.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2021). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. Proceedings of the IEEE, 107(3), 502–504.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W. M. Ramsey (Eds.), The Cambridge Handbook of Artificial Intelligence. Cambridge University Press.

Buolamwini, J. and Gebru, T., 2018, January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77–91) PMLR.

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B. and Barro, S., 2023. AI literacy in K-12: a systematic literature review. International Journal of STEM Education, 10(1), p. 29.

Durán, J. M. and Jongsma, K. R., 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics, 47(5), pp. 329–335.

Engadget. (2008). Bartle calls Blizzard out on torture quest in Wrath of the Lich King. Retrieved from https://www.engadget.com/2008-11-25-bartle-calls-blizzard-out-on-torture-quest-in-wrath-of-the-lich.html.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics. SAGE Publications.

Hovy, D. and Spruit, S. L., 2016, August. The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 591–598).

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 327 (October 2021), 39 pages. DOI: 10.1145/3476068.

Pastin, M. (2018). The Hard Problems of Management: Gaining the Ethics Edge. Jossey-Bass.

Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Pearson.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

Tyler, N. S., & Jacobs, P. G. (2020). The Role of AI and Machine Learning in Type 1 Diabetes: A Review. Diabetes Technology & Therapeutics, 22(11), 809–817.

Vasey, J., Havard, D., Holt, A., & Sarker, S. K. (2022). DECIDE-AI: New Reporting Guidelines to Bridge the Development to Implementation Gap in Clinical Artificial Intelligence. The Lancet Digital Health, 4(1), e13–e15.

Von Eschenbach, W. J., 2021. Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), pp. 1607–1622.