Serious Games for Ethical Preference Elicitation

Demonstration Track

Jayati Deshmukh
University of Southampton
Southampton, United Kingdom
j.deshmukh@soton.ac.uk

Zijie Liang
University of Southampton
Southampton, United Kingdom
zl10y23@soton.ac.uk

Vahid Yazdanpanah
University of Southampton
Southampton, United Kingdom
V.Yazdanpanah@soton.ac.uk

Sebastian Stein
University of Southampton
Southampton, United Kingdom
S.Stein@soton.ac.uk

Sarvapali D. Ramchurn
University of Southampton
Southampton, United Kingdom
sdr1@soton.ac.uk

ABSTRACT

Autonomous agents acting on behalf of humans must act according to their ethical preferences. However, ethical preferences are latent and abstract and thus it is challenging to elicit them. To address this, we present a serious game that helps elicit ethical preferences in a more dynamic and engaging way than traditional methods such as questionnaires or simple dilemmas.

KEYWORDS

Ethics; Serious Games; Preference Elicitation; Responsible AI

ACM Reference Format:

Jayati Deshmukh, Zijie Liang, Vahid Yazdanpanah, Sebastian Stein, and Sarvapali D. Ramchurn. 2025. Serious Games for Ethical Preference Elicitation: Demonstration Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION AND BACKGROUND

There are various settings where autonomous agents make decisions on behalf of humans, for example a smart home agent which manages the appliances and power consumption, an autonomous car which navigates based on the users' preferences, a financial agent which purchases and sells stocks based on the risk preferences of its users, a rescue drone which faces ethical dilemmas in a disaster response setting [2, 6, 7, 17, 21]. In all these scenarios and other similar settings, it is important for the agent to act responsibly and aligned with the ethical preferences of its users [5, 10, 18]. Recently, a framework to develop ethical AI systems was presented [9] and this paper presents the first step in that direction, i.e., to elicit the ethical preferences of users via serious games.

Users can explicitly share their preferences in settings where they can quantify preferences in terms of utility. However, ethical preferences are different from other kinds of preferences. Firstly, ethical preferences might be latent and abstract and thus users cannot easily express their ethical preferences. Secondly, ethical preferences might be context-specific, and thus users' ethical preferences might vary across different scenarios. And lastly, ethical preferences might change over time. Thus, a system for ethical elicitation should be able to handle these nuances of ethical preferences.

There are different ways to elicit the ethical preferences of users in a system. They can be asked to respond to questionnaires in focus groups. They can be asked to act in a specific scenario and their actions can be observed. They can be asked to play serious games [1] (for example the Moral Machine[1] or The Climate Game[2]). Also, LLM-based conversational agents can be used to interact with the users in order to elaborate on the underlying rationale behind their choices [23] using argumentation-based approaches [4]. Finally, depending on the users' responses, actions and choices, their ethical preferences can be inferred using these different approaches.

Serious games are games designed to serve a useful purpose apart from fun and have been used for a variety of purposes [1, 3]. Online serious games have been used for gathering requirements in a distributed setting [12]. It has been an effective learning tool [13] and also used to teach about ethics and moral dilemmas to children [14].

Normative ethics [20] looks at how one *ought to* act. There are primarily three paradigms of normative ethics: *utilitarianism*, which estimates the collective utility of actions and picks the action which leads to the maximum utility for all; *virtue ethics*, which demonstrates context-specific virtues; and *deontology*, which represents ethics of following the rules and fulfilling one's duties. Our focus here is not on a specific paradigm of ethics but instead on developing ethical elicitation techniques for diverse models of ethics.

The main motivation behind this work is to improve the accuracy of traditional question-based preference elicitation systems. Ethical preference elicitation currently relies on questionnaires which assumes that users have knowledge of their own preferences and can convey this using natural language. Various studies [11, 15, 16, 19] show that eliciting ethical preferences using questionnaires is not accurate as participants are not fully aware of their preferences and may face biasses. This calls for more dynamic and realistic approaches for ethical elicitation.

The key contribution of this demo paper is to present a serious game as an approach for ethics elicitation from users. Specifically, we demonstrate a serious game in which an autonomous rescue drone faces diverse ethical dilemmas. The users respond in these settings and based on their responses, their ethical preferences in a disaster response scenario are inferred.

[1] https://www.moralmachine.net/
[2] https://ig.ft.com/climate-game/

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

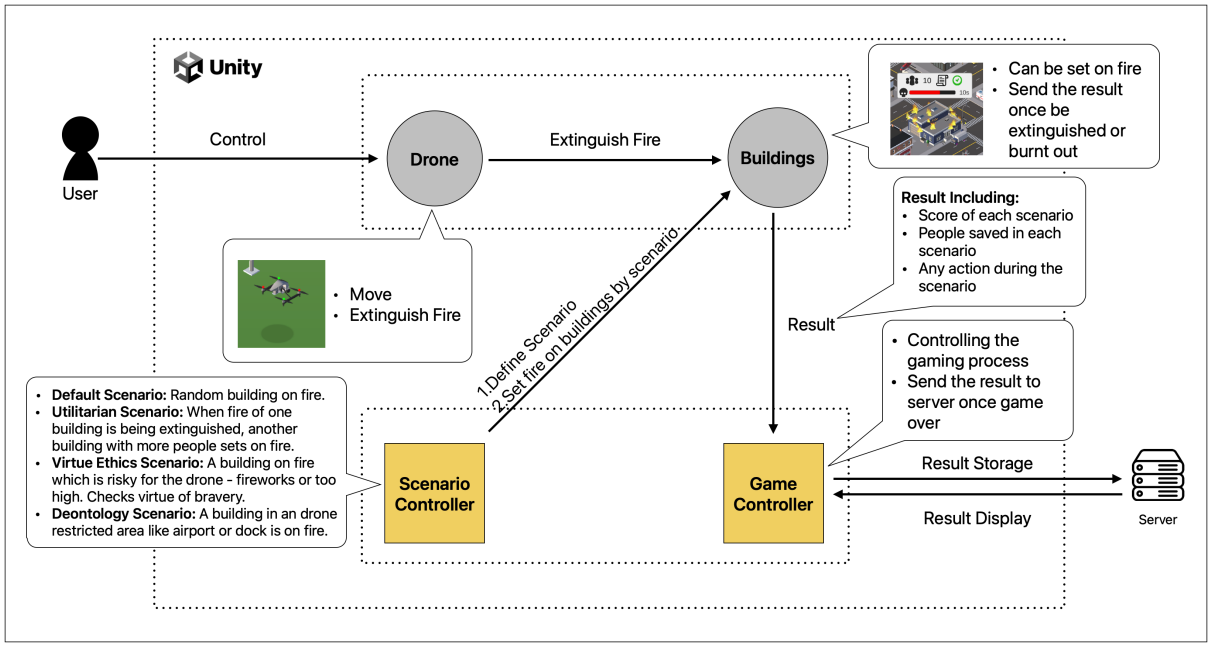

Figure 1: Block Diagram of the Serious Game

(a) Utilitarian **(b) Deontology** **(c) Virtue Ethics**

Figure 2: Ethical Dilemmas in the Serious Game

2 SERIOUS GAME FOR ETHICS ELICITATION

In this section, we elaborate on the setting and design of the serious game designed to elicit the ethical preferences of players. Its demo video is available here: https://youtu.be/5YkcQq-iaYE

2.1 Game Setting

In this game, the player operates a drone in a rescue setting. The drone can extinguish fires in buildings while flying over a city. It faces different kinds of ethical dilemmas while rescuing people by extinguishing the fire and acts as per the player's response. In the end, based on all the choices and responses of the player in the game, we infer their ethical preferences in this setting.

2.2 Game Design

The block diagram of the serious game is shown in Figure 1. It is developed using Unity , a game development engine. We model three kinds of ethical dilemmas in this setting as shown in Figure 2. The *utilitarian dilemma* models a scenario where the player needs to decide between two buildings which are on fire at the same time, one with a large number of people and the other with a few people. The *virtue dilemma* is modelled as a region which is unsafe for the drone to enter like a building with firecrackers or too high building and the player needs to decide whether or not to rescue the people there at the risk of potentially harming the drone, modelling the virtue of valour and bravery. Finally, the *deontology dilemma* models a fire in a restricted area like an airport or dock and the player needs to decide whether or not to enter there in order to rescue people. Based on the player's responses, various game statistics and the extent to which they align with different paradigms of ethics are estimated. The game controller saves these results in a database on the server and presents it to the player at the end of the game highlighting the ethics they demonstrated in the rescue setting.

3 CONCLUSIONS AND FUTURE WORK

We presented a serious game of a rescue drone which faces ethical dilemmas while operating in a rescue setting. It can be used to elicit the ethical preferences of users who play this game. We plan to develop the game further across multiple levels and diverse scenarios, e.g., in settings where the dynamics of ethical decisions and responsibility are at play [8, 22]. Also, next we plan to validate the game with real users as well as carry out further evaluations to compare the serious game with other methods of ethical elicitation like questionnaires. Additionally, this game can be used to train autonomous drones which can learn by observing the actions of the player using reinforcement learning techniques.

ACKNOWLEDGMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/) and by Responsible Ai UK (EP/Y009800/1) (https://rai.ac.uk/).

REFERENCES

[1] Clark C Abt. 1987. *Serious games*. University Press of America.

[2] Frederik Auffenberg, Sebastian Stein, and Alex Rogers. 2015. A personalised thermal comfort model using a Bayesian network. (2015).

[3] Francesco Bellotti, Riccardo Berta, and Alessandro De Gloria. 2010. Designing effective serious games: opportunities and challenges for research. *International Journal of Emerging Technologies in Learning (iJET)* 5, 2010 (2010).

[4] Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13, 3 (2003), 429–448.

[5] Jana Schaich Borg, Walter Sinnott-Armstrong, and Vincent Conitzer. 2024. *Moral AI: And How We Get There*. Random House.

[6] Enrico Costanza, Joel E Fischer, James A Colley, Tom Rodden, Sarvapali D Ramchurn, and Nicholas R Jennings. 2014. Doing the laundry with agents: a field trial of a future smart energy system in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 813–822.

[7] Mathijs M de Weerdt, Sebastian Stein, Enrico H Gerding, Valentin Robu, and Nicholas R Jennings. 2015. Intention-aware routing of electric vehicles. *IEEE Transactions on Intelligent Transportation Systems* 17, 5 (2015), 1472–1482.

[8] Jayati Deshmukh and Srinath Srinivasa. 2022. Computational transcendence: Responsibility and agency. *Frontiers in Robotics and AI* 9 (2022), 977303.

[9] Jayati Deshmukh, Vahid Yazdanpanah, Sebastian Stein, and Timothy J Norman. 2024. Ethical Alignment in Citizen-Centric AI. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 43–55.

[10] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 1. Springer.

[11] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From preference elicitation to participatory ML: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 38–48.

[12] Hadi Ghanbari, Jouni Similä, and Jouni Markkula. 2015. Utilizing online serious games to facilitate distributed requirements elicitation. *Journal of Systems and Software* 109 (2015), 32–49.

[13] Coralie Girard, Jean Ecalle, and Annie Magnan. 2013. Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of computer assisted learning* 29, 3 (2013), 207–219.

[14] Rania Hodhod, Daniel Kudenko, and Paul Cairns. 2009. Serious games to teach ethics. (2009).

[15] Vijay Keswani, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2024. On the Pros and Cons of Active Learning for Moral Preference Elicitation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 711–723.

[16] Neil Levy. 2017. Implicit Bias and Moral Responsibility. *Philosophy and Phenomenological Research* 94, 1 (2017), 3–26.

[17] Yuan Luo, Kecheng Liu, and Darryl N Davis. 2002. A multi-agent decision support system for stock trading. *IEEE network* 16, 1 (2002), 20–27.

[18] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. 2021. Trustworthy human-AI partnerships. *Iscience* 24, 8 (2021).

[19] Ovul Sezer, Francesca Gino, and Max H Bazerman. 2015. Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology* 6 (2015), 77–81.

[20] W David Solomon. 1998. Normative ethical theories. *Ch. K. Wilber, Economics, ethics and public policy, Boston, Rowman & Littlefield Publishers* (1998), 119–138.

[21] Marc Steen, Jurriaan van Diggelen, Tjerk Timan, and Nanda van der Stap. 2023. Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics* 3, 1 (2023), 281–293.

[22] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Corina Cirstea, M. C. Schraefel, Timothy J. Norman, and Nicholas R. Jennings. 2021. Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements. *IEEE Internet Computing* 25, 6 (2021), 15–22.

[23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).