A Brief Overview of an Approach Towards Ethical Decision-Making

Mashal Afzal Memon(✉)

Università degli Studi dell'Aquila, L'Aquila, Italy
mashalafzal.memon@graduate.univaq.it

Abstract. Ethics in decision-making reflects traits such as transparency, equity, and trust. However, when considering ethics in the decision-making process of autonomous agents, the significant challenge is how autonomous agents should interact to reach an agreement, knowing that their ethical preferences may differ. On that account, this study explores two fields to propose an approach to ethical decision-making: *automated negotiation*, the field concerning interaction among multiple agents to reach an agreement, and *machine ethics*, the field concerned with adding or ensuring moral behaviors from agents. Although agents can negotiate and decide on a solution automatically, whether they can propose an ethically correct decision is still a subject matter. To this end, this study proposes the concept of introducing ethics in the decision-making process of intelligent agents for ethical decision-making. In particular, we propose a research framework that addresses how user ethical preferences can be converted into quantifiable measures and further used by autonomous agents during negotiation for ethical decision-making.

Keywords: Adaptation and Learning · Automated negotiation · Ethical behavior of multi-agent systems

1 Introduction

Artificial intelligence has played an essential role in the development of future generation of intelligent agents capable of autonomous decision making [11,30]. Although the next generation of intelligent agents promises many advantages, their increased degree of freedom raises concerns about their moral behavior during decision-making [4]. Since the early 2000s, Picard emphasized the need for embedding morality into autonomous machines: "*the greater the freedom of a machine, the more it will need moral standards*" [28]. Consequently, the development of autonomous systems that can ensure the morality of their behavior has attracted the interest of the research community, leading to the birth of the field of "*Machine ethics*" [16]. When considering ethics in the decision-making process, a significant challenge is how autonomous agents should interact in order to reach a situational agreement, knowing that their ethical preferences may differ in general.

V. Malvone and A. Murano (Eds.): EUMAS 2023, LNAI 14282, pp. 458–464, 2023.
https://doi.org/10.1007/978-3-031-43264-4_32

Negotiation is a process between multiple agents in which a decision is made jointly by communication, i.e., through exchange of dialogues, bids, and offers to reach an agreement that is accepted by all agents [8,32]. In the context of "*Automated negotiation*", designing agents capable of effectively acquiring and integrating user ethical preferences into the decision-making process is a key challenge [6,13,21]. To this end, we focus on combining ethics with automated negotiation to propose an approach where autonomous agents negotiate with each other based on user ethical preferences for ethical decision-making. In particular, in this study, we propose a research framework that focuses on describing how user ethical preferences can be converted into quantifiable measures to be then used by autonomous agents during negotiation for ethical decision-making.

The remainder of the paper is structured as follows. In Sect. 2, we detail the related work to our study. Section 3 describes our research framework. Section 4 provides a discussion with an overview of the future research direction and Sect. 5 concludes the proposed study.

2 Related Work

In this section, we discuss related work that covers both the theoretical fundamentals and the current state-of-the-art for automated negotiation and ethical decision-making.

Challenges of introducing ethics in automated decision-making – In the following, we discuss more theoretical works that highlight the difficulties of introducing ethics into autonomous systems. In [26], Moor defines four different levels of ethical agents. At the lower levels, agents do not have any ethics explicitly added to their software, but may have an ethical impact on other agents, humans, and the environment due to their actions or design (e.g., autopilots can impact the safety of passengers). At higher levels, Moor identifies *explicit ethical agents*, who use available ethical knowledge in their decision process. At the higher level, Moor also introduces the concept of *fully ethical agents*, which are capable of making explicit judgments and justifying them, i.e., human-like ethical reasoning. In addition to the technical challenges related to the development of these agents [10,15,29], the uncertainty of different moral principles makes it difficult to identify a single ethical theory that can be followed to develop such intelligent systems [4,9,27]. In [14], Floridi describes digital ethics as two separate components. The first, *hard ethics*, represents the ethical rules described by the higher authorities, which are (in principle, should be) commonly accepted. The second, *soft ethics*, encompasses user morals, which can reflect on user personal preferences during decision-making. This vision poses the challenge of how to embed user ethical preferences into decision-making, not only in those situations in which humans interact with autonomous agents but also when the latter interact between themselves on behalf of humans.

Automated negotiation – In the following, we discuss studies that consider negotiation between autonomous agents for automated decision-making. In multi-agent systems, rules have been an effective technique for modeling negotiation. In [23], the seller and buyer agents negotiate the price using fuzzy rules to find the best bidding strategy. Agents learn during negotiation by interacting with opponents to modify and create new rules. A similar idea has been discussed in [18], where agents use associative rules during negotiation to adjust parameters such as time, value intervals to offer, and negotiation issues to reduce the number of interactions by generating associative rules. Although these studies present negotiation based on rules, negotiation based on user ethical preferences is still unexplored. Furthermore, the study in [20] defines various stages of the life cycle of the negotiating agent, and the studies in [7,25,31], propose multiple approaches to automate the negotiation process based on different stages of the life cycle of the negotiating agents. However, none of these studies considers user ethical preferences in negotiation, which is the main focus of our approach.

Ethical decision-making – We discuss below the studies that consider ethics during the decision-making process. The architecture of an artificial moral agent is proposed in [22], which combines the moral values of different stakeholders to make an ethical decision. The agent makes a decision by forming a single ethical theory from different moral values. It is assumed that moral values are classified and that agents utilize them to take a collaborative decision that leads to an agreement. However, in our study, where agents' moral values differ from each other and are unknown to opponents, rather than forming a single conclusion to agree, agents self-adapt their behavior and negotiate to reach an agreement until it satisfies their moral values. In [12], the study proposes an ethical reasoner to conduct decision-making. In this work, the ethical reasoner follows a predetermined ethical theory, and the possible actions that the system can undertake are ranked according to their adherence to the ethical theory. However, the proposed study does not consider the morality of users as part of decision-making, as the ethical principles followed by the system are decided by the system designers. Therefore, in our study, we focus on user morals for ethical decision-making instead of explicit ethical theories and rules.

3 Research Framework

This section describes the focus areas and research questions that result from the state-of-the-art. Figure 1 shows a visual representation of our framework.

Automated negotiation is a compelling research field that groups three familiar research fields into one, namely, game theory, economics, and artificial intelligence [5]. The significance of automated negotiation is receiving great attention in the current age, as intelligent agents that negotiate with each other and represent human users are likely to be more efficient [13,21]. On the other hand, machine ethics is a field that combines computational logic with moral philosophy [4]. A well-known obstacle in this field is the lack of general agreement on which specific ethical values should be followed by autonomous decision-making agents [10,15], as individuals differ in their moral judgements [4,27].

Traditionally, in a multi-agent environment, agents can be cooperative and communicate with each other to perform a shared task [13,21], or they can be selfish and compete with others to maximize their own utility [5,17]. In the

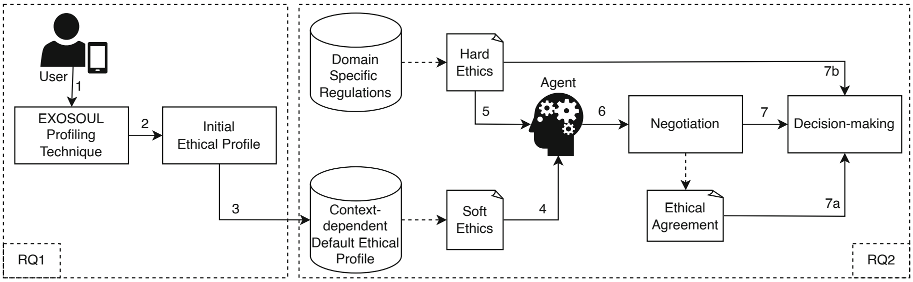

Fig. 1. A visual representation of the research framework. The dotted boxes highlight the elements dedicated to each research question. The solid box represents the component, the knowledge base represents the rules and ethical profile, the file icon represents the instance, the solid arrow represents the data flow, and the dotted arrow represents the connection between the instance and the parent element.

former case, the system can only follow the ethical principles decided by the system designers (thus disregarding plurality of opinions); whereas, in the latter case, the selfish behavior of agents will purposely lead them to ignore the ethical principles of others, as to maximize their own benefits according to their own ethical beliefs. For that, we employ the concept of ethics as proposed by Floridi [14], according to which soft ethics encompasses user ethical preferences and hard ethics represents explicit ethical rules. In our work, mimicking human behavior, we consider autonomous systems as independent and competing over shared resources but willing to negotiate to reach an agreement as long as it does not violate their own ethical boundaries. Note that it does not make much sense to hope that the agreement is reached once and for all; rather, it is situational in that it relates to or depends on specific circumstances, state of affairs, or environments. To this end, the research questions that model this study are:

RQ1: How does human ethical preferences can be represented as quantifiable measures?

This research question focuses on profiling human according to their ethical preferences (i.e., soft ethics as mentioned above). To reflect human ethical preferences in the decision-making process of autonomous agents, it is important to represent them as quantifiable measures. For that reason, within the EXOSOUL project[1] [1–3, 19], we exploit a personalized ethical profiling technique to collect individual's preferences through a survey (1) that aims to gather data on the moral preferences of users in the digital world. The resultant profile of this survey (2) is then used to develop a context-dependent ethical profile (3) that the autonomous agent uses for negotiation purposes, as shown in Fig. 1.

RQ2: How can we design autonomous agents that take human ethical preferences into account when negotiating for decision making?

[1] https://exosoul.disim.univaq.it/.

This research question focuses on detailing the process of combining human ethical preferences with automated negotiation. For that reason, in [24], we propose an approach in which an autonomous system adapts its behavior and adjusts its autonomy according to the input it receives from the user as an ethical profile (4). We assume to create a context-dependent profile from the general profile obtained through [1,2]. It is worth mentioning that even agents negotiate on the basis of soft ethics; we consider that each agent involved in the process is in compliance with domain specific rules (i.e. hard ethics as mentioned above) to avoid illegal actions (5). According to user ethical preferences, when the user shows priority towards herself, the agent self-adapts and becomes self-interested, and hence negotiates (6) to reach an agreement until it satisfies its ethical beliefs; however, if self-prioritization according to user ethical preferences is not important, the agent becomes cooperative and coordinates to reach an agreement if the opponent offers satisfy its ethical preferences. During negotiation, each received offer is then evaluated according to the ethical principles of the user profile. The negotiation ends when an ethical agreement is reached or no offer satisfies the ethical beliefs of the involved parties. When no agreement is reached, we consider the agents to follow domain-specific rules to apply a fall-back strategy for decision-making (7).

4 Discussion

This section provides a discussion and an overview of future steps. Our work highlights the need to consider ethics in the decision-making process of autonomous systems. This will help to ensure that autonomous systems behave ethically while enabling effective decision-making. To this end, as a first step, we propose to ingrain the ethical beliefs of the user into the system through an ethical profile [1–3,19]. Context-dependent ethical profiling is one of the future research directions of this work. For ethics-based negotiation, in [24], we then propose an approach to utilize the context-dependent ethical profile during negotiation. To this end, we consider the adoption of reinforcement learning as an appropriate technique. By employing reinforcement learning in negotiation, the agent will engage in a continuous loop to learn through user ethical preferences and adapt its negotiation strategy.

5 Conclusion

This study introduces an ethical perspective in the decision-making process of autonomous agents for ethical decision-making and details how an autonomous agent can represent user ethical preferences during negotiation. Negotiation resolves possible conflicts and results in ethical decisions that satisfy the user's ethical beliefs. In the future, we plan to implement this study and validate its effectiveness in real-world scenarios.

Acknowledgements. The authors would like to thank the entire multidisciplinary team of the EXOSOUL@univaq project for enlightening debates and joint work on digital ethics for autonomous systems.

References

1. Alfieri, C., Donati, D., Gozzano, S., Greco, L., Segala, M.: Ethical preferences in the digital world: the EXOSOUL questionnaire. In: HHAI 2023: Augmenting Human Intellect, pp. 290–299. IOS Press (2023)
2. Alfieri, C., Inverardi, P., Migliarini, P., Palmiero, M.: Exosoul: ethical profiling in the digital world. In: HHAI2022: Augmenting Human Intellect, pp. 128–142. IOS Press (2022)
3. Autili, M., Ruscio, D.D., Inverardi, P., Pelliccione, P., Tivoli, M.: A software exoskeleton to protect and support citizen's ethics and privacy in the digital world. IEEE Access **7**, 62011–62021 (2019). https://doi.org/10.1109/ACCESS.2019.2916203
4. Awad, E., et al.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)
5. Baarslag, T., Hendrikx, M.J., Hindriks, K.V., Jonker, C.M.: Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. Auton. Agents Multi-Agent Syst. **30**(5), 849–898 (2016)
6. Baarslag, T., Kaisers, M.: The value of information in automated negotiation: a decision model for eliciting user preferences. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, pp. 391–400 (2017)
7. Bachrach, Y., et al.: Negotiating team formation using deep reinforcement learning. Artif. Intell. **288**, 103356 (2020)
8. Bagga, P., Paoletti, N., Stathis, K.: Deep learnable strategy templates for multi-issue bilateral negotiation. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 1533–1535 (2022)
9. Bogosian, K.: Implementation of moral uncertainty in intelligent machines. Minds Mach. **27**(4), 591–608 (2017)
10. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. In: Artificial Intelligence Safety and Security, pp. 57–69. Chapman and Hall/CRC (2018)
11. Buiten, M.C.: Towards intelligent regulation of artificial intelligence. Eur. J. Risk Regul. **10**(1), 41–59 (2019)
12. Cardoso, R.C., Ferrando, A., Dennis, L.A., Fisher, M.: Implementing ethical governors in BDI. In: Alechina, N., Baldoni, M., Logan, B. (eds.) EMAS 2021. Lecture Notes in Computer Science, vol. 13190, pp. 22–41. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-97457-2_2
13. Chen, S., Weiss, G.: Automated negotiation: an efficient approach to interaction among agents. In: Interactions in Multiagent Systems, pp. 149–177. World Scientific (2019)
14. Floridi, L.: Soft ethics and the governance of the digital. Philos. Technol. **31**(1), 1–8 (2018)
15. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. **1**(6), 261–262 (2019)
16. Guarini, M.: Introduction: machine ethics and the ethics of building intelligent machines. Topoi **32**(2), 213–215 (2013)

17. Hoen, P.J., Tuyls, K., Panait, L., Luke, S., La Poutré, J.A.: An overview of cooperative and competitive multiagent learning. In: Tuyls, K., Hoen, P.J., Verbeeck, K., Sen, S. (eds.) LAMAS 2005. LNCS (LNAI), vol. 3898, pp. 1–46. Springer, Heidelberg (2006). https://doi.org/10.1007/11691839_1
18. Hu, J., Deng, L.: An association rule-based bilateral multi-issue negotiation model. In: 2011 Fourth International Symposium on Computational Intelligence and Design, vol. 2, pp. 234–237. IEEE (2011)
19. Inverardi, P., Palmiero, M., Pelliccione, P., Tivoli, M.: Ethical-aware autonomous systems from a social psychological lens. In: Proceedings of the 6th International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? CEUR Workshop Proceedings, vol. 3136, pp. 43–48 (2022)
20. Kiruthika, U., Somasundaram, T.S., Raja, S.: Lifecycle model of a negotiation agent: a survey of automated negotiation techniques. Group Decis. Negot. **29**(6), 1239–1262 (2020)
21. Kraus, S.: Agents that negotiate proficiently with people. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 137–137. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19656-0_21
22. Liao, B., Slavkovik, M., van der Torre, L.: Building jiminy cricket: an architecture for moral agreements among stakeholders. In: AAAI Conference on AI, Ethics, and Society, pp. 147–153 (2019)
23. Mahan, F., Isazadeh, A., Khanli, L.M.: Using an active fuzzy ECA rule-based negotiation agent in e-commerce. Int. J. Electr. Commer. Stud. **2**(2), 127–148 (2011)
24. Memon, M.A., Scoccia, G.L., Inverardi, P., Autili, M.: Don't you agree with my ethics? let's negotiate! In: Augmenting Human Intellect - Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI). Frontiers in Artificial Intelligence and Applications, vol. 368, pp. 385–388. IOS Press (2023)
25. Mohammadi Ashnani, F., Movahedi, Z., Fouladi, K.: Modeling opponent strategy in multi-issue bilateral automated negotiation using machine learning. Int. J. Web Res. **3**(2), 16–25 (2020)
26. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006)
27. Nallur, V., Collier, R.: Ethics by agreement in multi-agent software systems. In: 14th International Conference on Software Technologies, Prague, Czech Republic, 26–28 July 2019, pp. 529–535. SCITEPRESS (2019)
28. Picard, R.W.: Affective Computing. MIT press, Cambridge (2000)
29. Ryan, M., Stahl, B.C.: Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J. Inf. Commun. Ethics Soc. **19**(1), 61–86 (2020)
30. Totschnig, W.: Fully autonomous AI. Sci. Eng. Ethics **26**(5), 2473–2485 (2020)
31. Wu, L., Chen, S., Gao, X., Zheng, Y., Hao, J.: Detecting and learning against unknown opponents for automated negotiations. In: Pham, D.N., Theeramunkong, T., Governatori, G., Liu, F. (eds.) PRICAI 2021. LNCS (LNAI), vol. 13033, pp. 17–31. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89370-5_2
32. Zuckerman, I., Rosenfeld, A., Kraus, S., Segal-Halevi, E.: Towards automated negotiation agents that use chat interfaces. In: The Sixth International Workshop on Agent-Based Complex Automated Negotiations (ACAN) (2013)