HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230105

Don't You Agree with My Ethics? Let's Negotiate!

Mashal Afzal MEMON [a], Gian Luca SCOCCIA [b], Paola INVERARDI [b] and Marco AUTILI [a]

[a] *University of L'Aquila, Italy*
[b] *Gran Sasso Science Institute, Italy*

Abstract. The rapid growth in autonomous technology has made it possible to develop intelligent systems that can think and act like humans and can self-govern. Such intelligent systems can make ethical decisions on behalf of humans by learning their ethical preferences. When considering ethics in the decision-making process of autonomous systems that represent humans for ethical decision-making, the main challenge is agreement on ethical principles, as each human has its own ethical beliefs. To address this challenge, we propose a hybrid approach that combines human ethical principles with automated negotiation to resolve conflicts between autonomous systems and reach an agreement that satisfies the ethical beliefs of all parties involved.

Keywords. Autonomous Systems, Machine Ethics, Automated Negotiation, Ethical Decision Making

Innovation in autonomous technology has paved the way for the future generation of intelligent autonomous systems [1,2]. Their increased level of independence [3,4] raises concerns about their moral behavior in decision-making [5], leading to the birth of the field of "*Machine ethics*" [6,7]. An evident obstacle in this field [8,9,10], is the lack of general agreement on which ethical values should be followed by autonomous decision-making systems, as individuals differ in their moral judgements [5,11,12]. Therefore, when considering ethics in the decision-making process, a notable challenge is how autonomous systems should interact to reach a situational agreement, knowing that their ethical preferences may generally differ.

Automated negotiation is one of the prospects for solving conflicts between autonomous systems [13,14,15]. In a multi-agent environment, agents can be selfish and compete to maximize their utility [16,17], leading them to avoid the ethical beliefs of others. To address this challenge, we propose a hybrid approach that combines human ethical principles with automated negotiation. Traditionally, negotiation is a process of communication through bids, dialogues, and offers to reach an agreement [18]. Hence, in our approach, the interacting agents can negotiate to reach an agreement that, in a given context, satisfies the ethical beliefs of all involved parties[1], i.e., an *ethical agreement*.

As **motivating example** we consider a parking lot in a hospital where two independent autonomous connected vehicles compete for the nearest parking space on behalf of their passengers while the respective passengers have an emergency. Each vehicle is con-

[1] Note that the agreement is not definitive; rather, it depends on specific circumstances or environments.

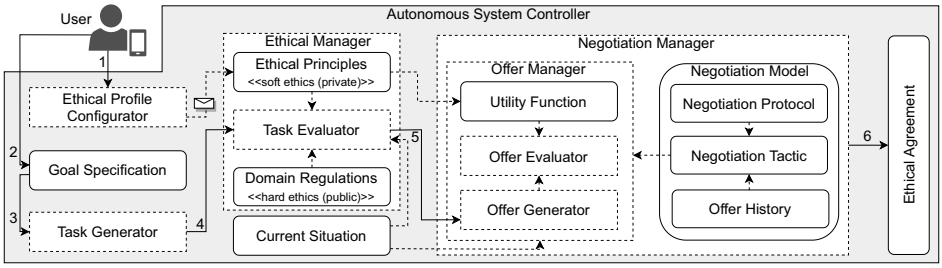

Figure 1. Overview of the proposed architecture (Dotted box = component; Rounded box = data; Solid arrow = data flow; Dotted arrow = dependency between sub-components).

figured with the ethical beliefs of the passenger, stored in an ethical profile. Each vehicle is aware of the urgency of its passenger to reach the hospital, but also of its willingness to negotiate to reach an agreement that satisfies, for all involved parties, their soft ethics (user ethical beliefs), while still complying with the hard ethics of the overall parking system (traffic laws).

Figure 1, shows an **architecture** to support the proposed approach. The user is in charge of uploading her ethical profile[2], e.g., by means of her mobile phone (1), and specifying the goal (2), i.e., drive to the nearest parking. From here, the autonomous system i.e., autonomous vehicle in our example, takes control and performs further simulation to achieve the goal [24]. Thus, the required tasks are generated by the system controller (3), and their outcomes are predicted via verifiable metrics [25]. Tasks are then evaluated against user ethical principles and domain-specific rules to measure their ethical impact in the current context to carry out actions that will achieve the goal (4). For this purpose, we employ the concept of ethics as proposed by Floridi [9,26], according to which *soft ethics* encompasses user ethical preferences, and *hard ethics* represents the ethical rules described by higher authorities, which are (in principle, should be) commonly accepted.

In a given context (e.g., the hospital parking), when resource contention is detected (e.g., competing for the same parking space), the Negotiation Manager is responsible for achieving a situational agreement (5). During the **negotiation**, offers are exchanged until an outcome is reached (ethical agreement or no agreement). Each offer specifies the tasks to be executed by the involved parties. Offers are generated (and evaluated) using a negotiation tactic and the current context. The utility of each proposed (and received) offer is computed based on user ethical principles, as each party has its own morals, and hence the results might differ. Each offer is then evaluated to determine whether to accept or reject it. We follow the intuition in [27], according to which ethical principles are considered soft constraints, rather than hard vetoes on tasks. Hence, our approach adjusts the autonomy so that, when no ethical option is available, the system strives to violate the (set of) least impactful ethical principle(s) (i.e., the *"least of all evils"*).

We define an ***ethical profile*** as $E_\phi = \{e_1 \succeq e_2 \succeq \cdots e_n\}$, where $e_1, \cdots, e_n$ are *ethical principles* sorted according to a total (not necessarily strict) order of importance $\succeq$ so that e_n is the least important (impactful) principle. Moreover, for negotiation purposes, we instantiate offers together with ethical principles into a context-dependent rule and define our ***ethical evaluation criteria*** as the formula: $accept(O^t) \Rightarrow_c max(E^t_\phi)$, with $E^t_\phi \subseteq E_\phi$.

[2]The profile is used by the controller to adjusts [19] the autonomy of the system. For this purpose, we will exploit the personalized ethical profiling technique (we are working on the multidisciplinary EXOSOUL project [20,21,22,23]), which accounts for the moral preferences of each individual user.

Following that criteria, the accepted offer O^t related to a task t in the context c maximizes the importance of the subset E^t_ϕ of ethical principles relevant to t. This means that, according to the chosen negotiation tactic, O^t maximizes the importance of those principles not violated among all possible offers that would be accepted by the negotiating counterpart. For instance, a "more qualitative" tactic may maximize the importance by giving priority to the most important principles. Another "more quantitative" tactic may instead maximize the importance by giving priority to the less important principles, thus preferring the number of not violated principles to their single importance. Still, another tactic would be to accept the violation of a principle only if all principles of minor importance are violated first.

Eventually, when an ethical agreement is reached (6), the system performs actions according to the tasks agreed upon by the negotiating parties. In our example, the vehicles will move to the parking they agreed upon. Alternately, if no ethical agreement is reached through negotiation, as no offer that satisfies the soft ethics of all involved parties could be found, the systems employ a fallback strategy for the decision-making, considering only the hard ethics of the current context.

The proposed hybrid combination of human ethical preferences with automated reasoning will help ensure that autonomous systems behave ethically while enabling effective decision-making. In the future, we plan to implement the proposed architecture and validate its performance in a real-world scenario.

References

[1] Werkhoven P, Kester L, Neerincx M. Telling autonomous systems what to do. In: Proceedings of the 36th European Conference on Cognitive Ergonomics; 2018. p. 1-8.

[2] Wang Y, Pitas I, Plataniotis KN, Regazzoni CS, Sadler BM, Roy-Chowdhury A, et al. On future development of autonomous systems: A report of the plenary panel at IEEE ICAS'21. In: 2021 IEEE international conference on autonomous systems (ICAS). IEEE; 2021. p. 1-9.

[3] Antsaklis PJ, Passino KM, Wang S. Towards intelligent autonomous control systems: Architecture and fundamental issues. Journal of Intelligent and Robotic Systems. 1989;1(4):315-42.

[4] Pratihar DK, Jain LC. Towards intelligent autonomous systems. In: Intelligent Autonomous Systems. Springer; 2010. p. 1-4.

[5] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature. 2018;563(7729):59-64.

[6] Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A. Implementations in machine ethics: A survey. ACM Computing Surveys (CSUR). 2020;53(6):1-38.

[7] Guarini M. Introduction: machine ethics and the ethics of building intelligent machines. Topoi. 2013;32(2):213-5.

[8] Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial intelligence safety and security. Chapman and Hall/CRC; 2018. p. 57-69.

[9] Floridi L. Establishing the rules for building trustworthy AI. Nature Machine Intelligence. 2019;1(6):261-2.

[10] Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society. 2020.

[11] Bogosian K. Implementation of moral uncertainty in intelligent machines. Minds and Machines. 2017;27(4):591-608.

[12] 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. SCITEPRESS; 2019. p. 529-35.

[13] Lopes F, Wooldridge M, Novais AQ. Negotiation among autonomous computational agents: principles, analysis and challenges. Artificial Intelligence Review. 2008;29(1):1-44.
[14] Kiruthika U, Somasundaram TS, Raja S. Lifecycle model of a negotiation agent: A survey of automated negotiation techniques. Group Decision and Negotiation. 2020;29(6):1239-62.
[15] Baarslag T, Hendrikx MJ, Hindriks KV, Jonker CM. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. Autonomous Agents and Multi-Agent Systems. 2016;30(5):849-98.
[16] Amir O, Sharon G, Stern R. Multi-agent pathfinding as a combinatorial auction. In: Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015. p. 2003-9.
[17] Hoen PJ, Tuyls K, Panait L, Luke S, La Poutre JA. An overview of cooperative and competitive multiagent learning. In: International Workshop on Learning and Adaption in Multi-Agent Systems. Springer; 2005. p. 1-46.
[18] 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. p. 6-10.
[19] Mostafa SA, Ahmad MS, Mustapha A. Adjustable autonomy: a systematic literature review. Artificial Intelligence Review. 2019;51:149-86.
[20] Autili M, Ruscio DD, Inverardi P, Pelliccione P, Tivoli M. A Software Exoskeleton to Protect and Support Citizen's Ethics and Privacy in the Digital World. IEEE Access. 2019;7:62011-21.
[21] 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?. vol. 3136 of CEUR Workshop Proceedings; 2022. p. 43-8.
[22] Alfieri C, Caroccia F, Inverardi P. AI Act and Individual Rights: A Juridical and Technical Perspective. In: Proceedings of the Workshop on Imagining the AI Landscape after the AI Act (IAIL 2022) co-located with 1st International Conference on Hybrid Human-Artificial Intelligence (HHAI'22). vol. 3221 of CEUR Workshop Proceedings; 2022. p. 43-55.
[23] Alfieri C, Inverardi P, Migliarini P, Palmiero M. Exosoul: Ethical Profiling in the Digital World. In: HHAI 2022: Augmenting Human Intellect - Proceedings of the 1st International Conference on Hybrid Human-Artificial Intelligence (HHAI'22). vol. 354 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2022. p. 128-42.
[24] Akkaladevi SC, Plasch M, Pichler A, Rinner B. Human Robot Collaboration to Reach a Common Goal in an Assembly Process. In: STAIRS; 2016. p. 3-14.
[25] Bremner P, Dennis LA, Fisher M, Winfield AF. On proactive, transparent, and verifiable ethical reasoning for robots. Proceedings of the IEEE. 2019;107(3):541-61.
[26] Floridi L. Soft ethics and the governance of the digital. Philosophy & Technology. 2018;31(1):1-8.
[27] Dennis L, Fisher M, Slavkovik M, Webster M. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems. 2016;77:1-14.