Explainability and self-disclosure for robot ethical introspection

Valeria Seidita*
Dipartimento di Ingegneria-Universitá degli Studi di Palermo
Italy
valeria.seidita@unipa.it

Antonio Chella†
Dipartimento di Ingegneria-Universitá degli Studi di Palermo
Italy ICAR-CNR, National Research Council, Palermo, Italy

ABSTRACT

Human-robot or human-AI interaction systems require a high degree of autonomy, proactivity, and adaptvity. The decisions that intelligent systems must make are highly dependent on the application context and trust is an essential element in task assignment. Explainability and ethical introspection capabilities are important in building trust and understanding in artificial processes. In this paper, we present our ongoing work aimed at equipping robots with ethical introspection capabilities when interacting with humans by designing and implementing explainable and self-disclosure capabilities. Using a computational model of ethical introspection that incorporates theories of psychology, ethics, and AI, we built robots that examine and reflect on their actions to evaluate and validate them. We use the Belief-Desire-Intention (BDI) agent paradigm and related programming languages along with the speech act mechanism to improve and extend the robot's ethical values to better guide its decision-making process and the impact it has on humans.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Ethics, Self-Disclosure, Ethical Introspection

ACM Reference Format:
Valeria Seidita and Antonio Chella. 2018. Explainability and self-disclosure for robot ethical introspection. In *Proceedings of 2nd International Workshop on Multidisciplinary Perspectives on Human-AI Team Trust (Multittrust 2.0).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Robots that emulate human behavior and assist individuals in their daily lives have long been a dream. A pioneer of this concept was Isaac Asimov, whose Three Laws of Robotics encapsulated the notion of robots interacting with humans. Asimov's three laws provide a valuable foundation; however, they are not sufficient when addressing contexts where robots must exhibit autonomy, proactivity, and self-adaptation.

The scenarios under consideration pertain to human-robot team interaction (HRTI) [11][5][6]. Within HRTI, the focal point is the collaboration between teams of humans and robots to achieve common goals. Irrespective of the nature of the task, whether it is purely social, such as providing companionship to an elderly individual, or high-risk, as in military settings, humans and robots must exchange information regarding the objectives, mission, their respective limitations, capabilities, and the work environment.

Typically, drawing upon their knowledge of the aforementioned elements, each team member selects an action to fulfill their mission. Nevertheless, this process is not strictly individualistic. The choice of action is also contingent upon the presence and actions of other team members, their knowledge, and competencies. Each team member must possess the ability to comprehend and anticipate the actions of their peers, as well as make decisions regarding which actions in the plan they can execute. Ultimately, each team member must decide which actions to undertake personally and which to delegate.

Several factors come into play in this decision-making process, with trust in one another being a pivotal element. One of the key factors that engenders or enhances trust in a fellow team member is the ability to expound on the rationale behind their actions. In a complex environment characterized by high dynamism, leading to uncertainty and decision-making challenges, such as in healthcare or military contexts, the selection of the optimal action is contingent upon the ability to evaluate the outcomes of actions in relation to predefined goals and conditions.

In some of our previous work, we have explored strategies for enhancing a robot's decision-making abilities through the concept of 'anticipation' [4][7][15]. We have developed a tool that allows the robot to transparently present its decision-making process. The robot selects an action given at design time and simulates the outcome before execution. If the simulated result aligns with the post-conditions of the goal, the mission is deemed successful; otherwise, the robot must opt for an alternative action. Concurrently, the robot provides an explanation to its human companion regarding its actions.

In any complex scenario involving the autonomy of a robot in the interactive domain, considerations concerning 'ethics' also come to the forefront. For instance, the potential invasion or violation of privacy when employing robots to assist patients necessitates that robots adhere to two fundamental principles: *(i)* making decisions aligned with the ethical standards of the society in which they operate, and *(ii)* articulating the rationale behind their actions to build human trust and influence human decisions. For example, in the context of a robot assisting a medical doctor, the robot

*Both authors contributed equally to this research.
†Both authors contributed equally to this research.

Multittrust 2.0, December 4, 2023, Gothenburg, Sweden

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

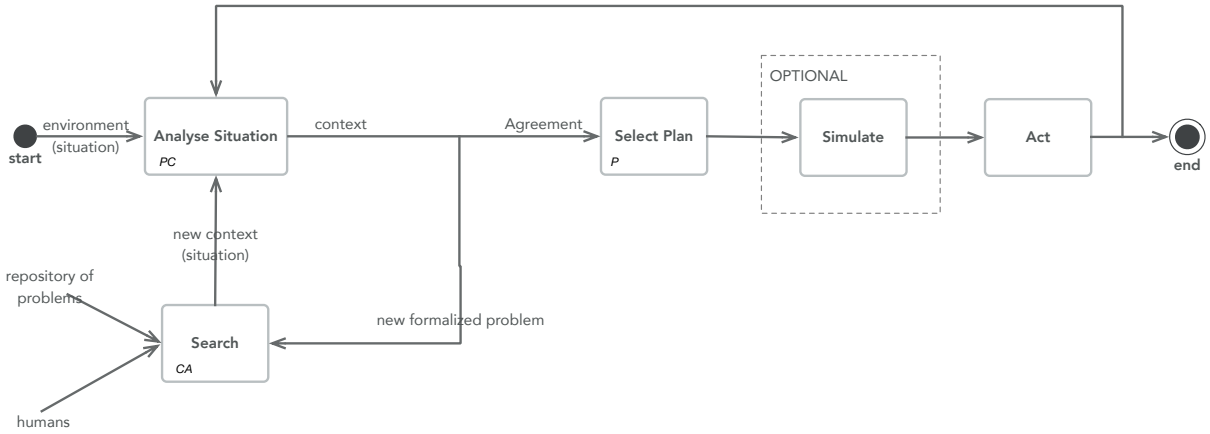

Figure 1: The reasoning process of a system equipped with Artificial Phronesis

can suggest an action to the doctor, justifying it from an ethical perspective, thereby stimulating thoughtful consideration. This process supports not only efficient decision-making but also provokes reflection in the human user.

Our endeavor revolves around the exploration of the roles of introspection and self-disclosure in ethical deliberation and social influence. In this paper, we present an initial hypothesis on how to formulate a computational model of ethical introspection in robots, building upon prior work on anticipation and trust [4][12][8].

2 ENABLING ETHICAL REASONING IN ARTIFICIAL AGENTS: A MODEL

Ethical introspection is an activity that focuses on mental states and mental processes as they occur or have just occurred. Our concept of endowing a robot with human-driven ethical introspection is to integrate it with the robot's self-awareness. Self-awareness implies a critical examination of one's own actions, intending to comprehend the ethical ramifications of those actions. In our approach, we draw inspiration from psychological, philosophical, and neuroscientific studies on ethical introspection, with Sullins serving as primary references [17][18][9]. Sullins proposed a theoretical framework based on the concept of 'artificial phronesis,' which pertains to human practical reasoning abilities and the virtue of rational thinking. 'Phronesis' represents philosophical wisdom in practical ethics and serves as the foundation for constructing machines capable of emulating human moral and practical reasoning.

Sullins asserts that ethical decision-making cannot always be reduced to a simple 'if situation (x), then action (y)' formula. Ethical decisions are predicated on the agent's habits, practices, and a comprehensive analysis of all elements of an ethical problem and potential consequences of actions. The framework proposed by Sullins involves two interacting agents: one for problem classification and one for case analysis. The former examines the situation to extract critical features, such as context and urgency. The latter collaborates in searching for analogous ethical cases within specific repositories and online sources to enhance the analysis.

Our objective is to fuse computational models of introspection, self-awareness, justification, and anticipation with mechanisms for ethical reasoning. In Figure 1, we present an integration of Sullins' model with our work spanning several years. In structuring the decision-making process, we also take internal states into account. To this end, we have conducted experiments and opted to employ BDI (Belief-Desire-Intention) agent technology [13][10], utilizing the JaCaMo implementation framework [2][3]. BDI agents aptly embody practical thinking, encompassing the process of agents determining how to translate their intentions into actionable steps within their environment.

Practical reasoning encompasses actions such as planning, resource allocation, and action sequence management, all of which consider agent constraints, environmental conditions, and potential consequences. This aligns well with the underlying concept of 'artificial phronesis.' Ethical reasoning can involve multiple agents, including humans. In Figure 1, we depict three artificial agents: the problem classification agent (PC), the case analysis agent (CA), and the planner agent (P). Upon encountering an ethical quandary, the first two agents engage in continuous interaction at the outset of the reasoning process. The initial step involves inspecting the environment and situation to formulate the problem and context. The CA leverages this context to seek analogous situations, potentially engaging with humans. A novel scenario is then created and analyzed, potentially prompting a reevaluation of the problem. Eventually, the CA and PC concur on the context and contact the planner, who selects a plan and executes actions, which may also be simulated. These actions impact the environment, initiating a new situation and reinitiating the cycle. Actions receive positive or negative feedback, signifying their contribution to or deviation from the desired goal. This cycle embodies a sensing-decision-action process, with key points focused on ethical introspection through self-disclosure.

Figure 2 illustrates the reasoning cycle of each agent of Figure 1, highlighting the points at which we introduce extensions to, in order: (i) generate a self-model, (ii) explain actions, and (iii) facilitate ethical introspection. Our goal is to adapt the reasoning cycle of the agents featured in Figure 1, as depicted in Figure 2. To create an agent with ethical introspection capabilities, we prioritize the need for explanations, even before it needs to construct a self-model.

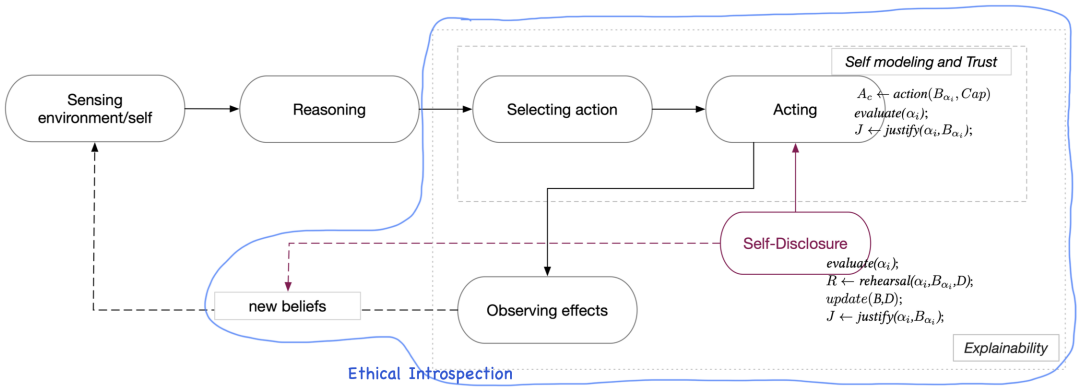

Figure 2: Towards the implementation of self-disclosure and ethical introspection. The reasoning cycle of each agent and the points in which we add functions at the interpreter level (low implementation level) for realizing self-modeling, explainability and introspection.

From a technical standpoint, actions can be explained as a function of the couple <belief, capabilities>. The agent decomposes the plan for mission execution into actions, closely associated with its knowledge of those actions and its ability to execute them. Consequently, the agent maintains a self-model and justifies action outcomes. We propose that, for each action, a rehearsal function can be defined, closely tied to self. The rehearsal function facilitates feedback with self, allowing the agent to reassess and adjust its knowledge base, beliefs, and decision-making process. This adjustment is based on observations of the effects of actions and new generated beliefs. When actions encompass ethical norms, the refined beliefs result from ethical introspection.

The practice of rehearsing has been implemented in previous works through the mechanism of speech acts [16]. A speech act [14][1] embodies the communicative action of the agent in which it expresses its actions, thus enabling self-revelation capabilities within an agent or agent system. The effect on the thinking process is indirect, indeed there is a change in the agents' knowledge as the result of speech acts. Speech acts along with agent technology allow to autonomously produce self-disclosure and then explainability.

3 CONCLUSION

This paper presents our ongoing research focused on the implementation of an ethical introspection model within robots endowed with self-awareness. Our research endeavors encompass computational models of introspection, self-awareness, justification, and anticipation, integrated with processes conducive to ethical deliberation. The principal objective underlying this research has been the facilitation of ethics within the decision-making framework of artificial agents. Initial progress was achieved through the integration of Sullins' model with our prior advancements, utilizing a Belief-Desire-Intention (BDI) agent-based paradigm. This model inherently captures the agent's practical reasoning process, placing significant emphasis on aspects such as planning, resource allocation, and the management of action sequences.

Furthermore, we have introduced an extended facet to the model, integrating ethical practice through the medium of speech acts. This augmentation empowers the agent to articulate and elucidate its actions. This communicative process catalyzes ethical evaluation and self-awareness, with the overarching aim of augmenting the decision-making capabilities and ethical conduct of these agents.

In the future, our research trajectory will involve further refinement of the model, drawing inspiration from in-depth interdisciplinary studies spanning the realms of psychology, philosophy, and neuroscience, all of which delve into the domains of ethics and introspection.

Collectively, our work lays the foundation for an in-depth exploration of ethics and introspection within the domain of autonomous robots, holding substantial potential for application in diverse domains, including healthcare and assistive robotics, among others. The incorporation of ethical models into robots represents a pivotal stride towards the responsible utilization of technology in complex and dynamic contexts, thereby enhancing the synergy between machines and human agents.

ACKNOWLEDGMENTS

International Exchanges 2022 Cost Share (Italy only) IEC\R2\222031 - Joint Research Program on Assistive Robots using Theory of Mind

REFERENCES

[1] John Langshaw Austin. 1975. *How to do things with words.* Vol. 88. Oxford university press.
[2] Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, and Andrea Santi. 2013. Multi-agent oriented programming with JaCaMo. *Science of Computer Programming* 78, 6 (2013), 747–761.
[3] Rafael H Bordini, Jomi Fred Hübner, and Michael Wooldridge. 2007. *Programming multi-agent systems in AgentSpeak using Jason.* John Wiley & Sons.
[4] Cristiano Castelfranchi, Antonio Chella, Rino Falcone, Francesco Lanza, and Valeria Seidita. [n. d.]. Endowing robots with self-modeling abilities for trustful human-robot interactions. *framework* 2 ([n. d.]), 1.
[5] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. 2017. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775* (2017).
[6] Antonio Chella, F.rancesco Lanza, and Valeria Seidita. 2018. A Cognitive Architecture for Human-Robot Teaming Interaction. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition.* Palermo.
[7] Antonio Chella, Francesco Lanza, and Valeria Seidita. 2019. Decision Process in Human-Agent Interaction: Extending Jason Reasoning Cycle. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence*

and Lecture Notes in Bioinformatics), Vol. 11375 LNAI. Springer, Cham, 320–339. https://doi.org/10.1007/978-3-030-25693-7_17

[8] Antonio Chella, Arianna Pipitone, Francesco Lanza, and Valeria Seidita. 2023. Toward Virtuous Machines: When Ethics Meets Robotics. In *Ethics in Research: Principles and Practical Considerations*. Springer, 81–91.

[9] Antonio Chella, Arianna Pipitone, and Jonh P. Sullins. 2024. Competent Moral Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis. In *Trolley Crash: Approaching Key Metrics for Ethical AI Practitioners, Researchers, and Policy Makers*.

[10] Lavindra De Silva, Felipe Rech Meneguzzi, and Brian Logan. 2020. BDI agent architectures: A survey. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020, Japão*.

[11] Lanssie Mingyue Ma, Terrence Fong, Mark J Micire, Yun Kyung Kim, and Karen Feigh. 2018. Human-robot teaming: Concepts and components for design. In *Field and Service Robotics: Results of the 11th International Conference*. Springer, 649–663.

[12] Arianna Pipitone, Alessandro Geraci, Antonella D' Amico, Valeria Seidita, and Antonio Chella. 2023. Robot' s Inner Speech Effects on Human Trust and Anthropomorphism. *International Journal of Social Robotics* (2023), 1–13.

[13] Anand S Rao, Michael P Georgeff, et al. 1995. BDI agents: from theory to practice.. In *Icmas*, Vol. 95. 312–319.

[14] J.R. Searle and John R. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge university press.

[15] Valeria Seidita, Cristina Diliberto, Paolo Zanardi, Antonio Chella, and Francesco Lanza. 2019. Inside the robot's mind during human-robot interaction. In *7th International Workshop on Artificial Intelligence and Cognition, AIC 2019*, Vol. 2483. CEUR-WS, 54–67.

[16] Valeria Seidita, Angelo Maria Pio Sabella, Francesco Lanza, and Antonio Chella. 2023. Agent talks about itself: an implementation using Jason, CArtAgO and Speech Acts. *Intelligenza Artificiale* 17, 1 (2023), 7–18.

[17] John P Sullins. 2019. The Role of Consciousness and Artificial Phronēsis in AI Ethical Reasoning.. In *AAAI spring symposium: towards conscious AI systems*.

[18] John P Sullins. 2021. Artificial Phronesis. *Science, Technology, and Virtues: Contemporary Perspectives* (2021), 136.