Responsible Autonomy for Hybrid Intelligence

Doctoral Consortium

Anastasia S. Apeiron
Utrecht University
Utrecht, Netherlands
a.s.apeiron@uu.nl

ABSTRACT

In hybrid intelligence (HI) systems, artificial intelligence (AI) agents and humans work together to solve complex tasks. In these interactions, each agent is expected to work autonomously and be responsible for their actions. By capturing consent as a regulation of actions in a normative environment (such as a HI system), an agent can determine an appropriate action within the normative environment, and reason on the moral and ethical requirements and effects of the action. Current consent representations do not allow agents to reason on normative actions, which limit agent autonomy. We are developing a representation of consent that captures the nuances of consent from human-human interaction, and expresses them computationally to allow the AI agent to responsibly practice autonomy in a HI system. In future work, the proposed representation will be evaluated against human intuitions about consent, and compared to current consent representations to ensure a robust and domain-agnostic formalisation. Further research includes developing a consent representation that can manage multi-party consent and shared resources, specifying accounts for consent violations to determine culpability, and exploring a developmental approach to norm representation and management for greater perceived agent responsibility and autonomy in a HI system.

KEYWORDS

Consent; Social Norms; Autonomy

ACM Reference Format:
Anastasia S. Apeiron. 2025. Responsible Autonomy for Hybrid Intelligence: Doctoral Consortium. 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

When an artificial intelligence (AI) agent interacts with a human in a hybrid intelligence (HI) system, the behaviour of the agent is expected to be responsibly autonomous [3, 10] where the rules surrounding the interaction can be represented, managed, and explained from an ethical and moral standpoint [2]. Consent is used in human interactions to regulate autonomy [8] and has normative power [5, 6], where a person can express the temporary and specific modification of norms to authorise the use of their resource by another person (which may be otherwise violating social norms and incurring sanctions). Consent captures the norms surrounding these interactions [7], and a computational representation of consent can enable the responsible autonomy of an AI agent in a HI system [8].

In HI systems, agents may need to take action on behalf of a human, manage access to data by other agents, and/or autonomously make decisions about delegating another agent's resources. In current consent representations, agents mainly refer to consent as a form of authorisation, and do not reason on when consent is required, from whom, and what impact the given consent has on the social norms already in the system. For these interactions, such as on websites, the General Data Protection Regulation (GDPR) protects the privacy of website users by allowing users to determine the amount of consent they wish to give [4]. The websites represent and manage the consent expressed by various users using Consent Management Providers (CMPs), which act like a bookkeeping service to lookup and share the relevant consent with the relevant parties when necessary [4]. While CMPs are mostly used for consent regarding what to do with user data within the narrow context of websites, a HI system where humans and AI agents work together to accomplish tasks requires a finer-grained representation and management of consent.

For example, in a HI system, a human can give consent to their agent to send invitations for meetings on their behalf, and also consent to their agent propagating this consent to other agents in the system so the other agents may forward these invitations to other agents; what happens when an agent who has the human's consent to invite others to meetings invites someone the human does not wish to invite? This example highlights an important question that drives this research: where does one consent end and another begin? Has the human waived their right to decide who can join the meetings?

Current consent representations are domain-specific and limit the autonomy of the agent in changing normative environments when the agent cannot adapt their reasoning on consent accordingly [9]. An example of this is an AI agent lending the car of one person to another person under strict conditions, but these conditions are not met due to extenuating circumstances after the car has been lent. This example raises another question, namely, how can an agent adapt their reasoning in a changing normative environment to avoid a consent violation? To answer these questions, we must first determine the constituents of consent, followed by how the interactions between these constituents function, and lastly, what the impact and outcomes of these interactions in the greater scope of a HI system are.

In this research, we approach our questions with a human-centred perspective by first grounding the consent representation in the relevant philosophical literature, followed by constructing

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).

a formal representation, and then computationally implementing and evaluating that representation with humans and AI agents.

2 CURRENT RESEARCH

In our recent work [1], we propose a consent model and a set of mechanisms that formally define consent and represent its lifecycle.

Consent is necessary when an agent requires access to resources that are not under their ownership, or sovereignty. A social norm protecting the sovereignty of an agent cannot be violated without normative consequences, and consent prevents the enactment of these consequences.

2.1 Consent Model

We define consent as the tuple containing two agents, a set of norms, a stated goal, and an action.

Firstly, an agent must deliberate on whether consent is necessary in a given situation and from whom to request it if so. If the resource the agent seeks is not under their sovereignty, consent is necessary and must be solicited from the agent that is the sovereign of the resource, which defines the two agents in the consent instance, namely the consent receiver and the consent giver.

Secondly, the norms contained within a consent instance must include at least one authorisation and one commitment to regulate the conditions under which a resource is accessible to another agent that is not the owner (authorisation), and to specify the outcomes of accessing the resource (commitment). These norms denote the beginning and end points of the consensual interaction; the action specified in the consent instance is authorised to allow the consensual access to a specific resource, and the consent receiver agent commits to bring about their stated goal once they have accessed the resource. By only specifying these two points, an agent has greater autonomy in adapting their proximal actions in a changing normative environment while also being held accountable for upholding their commitments.

2.2 Consent Lifecycle

An active consent instates a negotiated set of norms into the socio-technical system (STS) to allow an agent to infringe upon a social norm; when is consent activated? Following the norm lifecycle, the consent lifecycle begins with the creation of the consent instance with two agents, an empty set of negotiated norms, a stated goal, and an action. A consent instance can be either solicited, where a set of norms are negotiated between the consent receiver and giver agents, or unsolicited, where the set of norms are not negotiated and instated outright by the resource-sovereign agent. Once the consent instance is active, the norms within the consent instance are also active. The consent instance terminates when (1) the expiration condition of the authorisation has been reached, (2) the agreement on the norms of the consent instance are withdrawn, (3) the commitment is fulfilled and the stated goal of the consent instance becomes true in the STS, or (4) any of the norms outlined within the consent instance are violated.

3 FUTURE DIRECTIONS

As a part of our future work, we aim to evaluate the impact of a human-inspired consent representation on normative interactions between two agents in a HI system. As a first step, we seek to evaluate the proposed consent representation in comparison to human intuitions on consent through a user study.

> **RQ 1:** How does a human-inspired consent representation affect the perceived responsibility and accountability of a HI system?

In this user study, we explore the effectiveness and usability of our human-inspired consent representation by comparing the outcomes of consensual interactions from the consent model to the expected outcomes of the interaction produced by the human interlocutor. Secondly, we aim to explore norm emergence and norm violation detection across various consent representations.

> **RQ 2:** How does a human-inspired consent representation affect norm emergence and norm violation detection in a HI system compared to current consent representations?

In this simulation, we survey dyadic consensual interactions between agents that are both using human-inspired consent representations, both using current consent representations, and one using the human-inspired and one using the current consent representation. Furthermore, we will explore the representation and management of multi-party consent, where consensual interactions include more than two agents.

> **RQ 3a:** What are the constituents of multi-party consent and how do they interact with each other?
> **RQ 3b:** How does the lifecycle of consent change between dyadic versus multi-party interactions involving consent?

Developing a consent representation that can manage consent in both dyadic and multi-party interactions promotes greater autonomy of the AI agent across different normative environments, and allows for resource sharing between multiple agents with, possibly, varying levels of sovereignty over the resource. An example of such a scenario is when an AI agent must decide whether to post a group photo online when each participant of the photo may have varying levels of consent. Moreover, we aim to explore blameworthiness and culpability in the event that there is a consent violation, where an agent's intention and choice of actions can be used to provide an account of the severity of the consent violation.

> **RQ 4a:** What are the levels of severity for a consent violation and how are they determined?
> **RQ 4b:** Does a system to determine culpability increase perceived responsibility of the HI system?

Determining the severity of a consent violation helps determine the appropriate sanctions that are applicable for the violating agent. By developing a robust and reliable way of providing an account for consent violations, the perceived responsibility of the agent may increase. Lastly, exploring the developmental psychology of norm awareness may provide a more human-like consent representation in artificial agents, especially if these cognitive process of human infants can be mimicked in artificial agents.

ACKNOWLEDGMENTS

This project is funded by the Hybrid Intelligence Centre, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (https://www.hybrid-intelligence-centre.nl/).

REFERENCES

[1] Anastasia S. Apeiron, Davide Dell'Anna, Pradeep K. Murukannaiah, and Pınar Yolum. 2025. Model and Mechanisms of Consent for Responsible Autonomy. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. Forthcoming.

[2] Virginia Dignum. 2017. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4698–4704.

[3] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S Kließ, Maite Lopez-Sanchez, et al. 2018. Ethics by design: Necessity or curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 60–66.

[4] Maximilian Hils, Daniel W Woods, and Rainer Böhme. 2020. Measuring the emergence of consent management on the web. In *Proceedings of the ACM Internet Measurement Conference*. 317–332.

[5] Heidi M Hurd. 2015. The normative force of consent. *Forthcoming in the Routledge Handbook on The Ethics of Consent, Peter Schaber ed.(Routledge Press, 2016)., University of Illinois College of Law Legal Studies Research Paper* 15-36 (2015).

[6] Neil C Manson. 2016. Permissive consent: a robust reason-changing account. *Philosophical Studies* 173 (2016), 3317–3334.

[7] Neil C Manson. 2022. Autonomy and Consent. In *The Routledge Handbook of Autonomy*. 357–367.

[8] Munindar P Singh. 2022. Consent as a foundation for responsible autonomy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12301–12306.

[9] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. 2014. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE transactions on knowledge and data engineering* 27, 1 (2014), 193–206.

[10] Jessica Woodgate and Nirav Ajmeri. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. *Comput. Surveys* (2024), 1–37.