Emergent Responsible Autonomy in Multi-Agent Systems

Doctoral Consortium

Jayati Deshmukh
International Institute of Information Technology, Bangalore
Bangalore, India
jayati.deshmukh@iiitb.org

ABSTRACT

Autonomous agents operating in multi-agent environments, face the *dilemma of responsibility* where they must choose between actions that are individually beneficial versus those that are considered responsible and ethical. Current approaches address this problem either using external reinforcements, or intrinsic notions of ethics that act as constraints overriding the agents' rational choice. Both of these approaches become difficult to scale across complex, dynamic situations, where the responsibility dilemma has to be resolved dynamically. Thus, there is a need to design models of agency, where a sense of ethics and responsibility are an integral part of the agent model and not in conflict with agents' self-interest dynamics. Towards this end, this thesis proposes a model called Computational Transcendence (CT) in which, an agent's *"sense of self"* is made elastic, that enables it to dynamically *identify* with external elements of its environment like other agents, communities and concepts. Agents continue to act rationally, working towards utility maximisation, but their utility is determined by their sense of self formed from their elastic identities. We show that this leads to emergence of responsible autonomy, in various multi-agent network conditions.

KEYWORDS

Responsible AI; Autonomy; Identity; Multi-Agent Systems

ACM Reference Format:
Jayati Deshmukh. 2023. Emergent Responsible Autonomy in Multi-Agent Systems: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Autonomous agents are getting prevalent in a variety of multi-agent scenarios [9]. They are usually designed to maximise their utility in the specific context they operate. However, these agents operate in a shared state space such that their actions impact other autonomous agents and humans in the system. Thus, there is a pressing need to design autonomous agents which act responsibly [2, 8, 10].

A variety of approaches and paradigms have been used to design ethical and responsible autonomous agents [1, 11]. Top-down approaches use norms and rules to ensure that agents act ethically. While bottom-up approaches use reinforcements and learning based mechanisms to design ethical agents. Also, there are hybrid techniques which combine both these approaches. Even different types of ethical paradigms like deontology, virtue ethics, utilitarianism etc. have been modelled in autonomous agents [8, 11]. However, it is complex and difficult to model all possible types of scenarios in advance in order to train agents to act ethically in every possible context. Also, in most existing approaches and paradigms, ethics and responsible behaviour are an after-thought, something which the agents should comply with, while primarily trying to achieve their goals.

We believe that responsible behaviour is not an *extra* feature, rather it is an intrinsic property of agents [4]. We can teach agents how to *act* responsibly or we can design agents which can *be* responsible. There are paradigmatic differences in these two types of autonomous agents and the way in which they perceive their surroundings, operate and make decisions. Also, intrinsically responsible agents need not be trained for each and every scenario, they can be responsible even in new, unseen contexts.

We present a model called Computational Transcendence (CT) [7], which modifies the identity of autonomous agents such that they can incorporate different aspects of the system (like other agents, collectives of agents and even systemic concepts) into their *sense of self*. Transcended agents are not told what they *ought to do*, rather who they *ought to be*. We show that this elastic identity of autonomous agents results in emergent responsible behaviour.

We have evaluated transcended agents in different scenarios such as 2-player games like Prisoners' Dilemma, 3-player games with collusion and Iterated Prisoners' Dilemma in a network of agents [7]. Recently, we have also demonstrated promising results of transcended agents in realistic scenarios like supply chains and traffic management [6] and also compared CT with other ethical paradigms [5]. Next we present the core CT model, discuss some of the key results and elaborate our future directions.

2 MODEL AND RESULTS

We present the core idea of transcendence in a multi-agent system with autonomous agents. Formally, a transcended agent, a has a sense of self, $S(a)$ which is modelled as follows:

$$S(a) = (I_a, d_a, \gamma_a) \tag{1}$$

Here, I_a represents the identity set of aspects like agents, collectives and concepts which the agent identifies with, d_a is the semantic distance of the agent which denotes the perceived logical distance of an agent to each aspect in its identity set and γ_a is the transcendence level of the agent which denotes the extent to which it identifies with others in the system. A low value of γ_a denotes a self-centred agent, while a high value denotes an agent which identifies with other aspects to a greater extent. An agent a, with transcendence level γ_a identifies with an aspect o whose distance is $d_a(o)$ with an attenuation factor of $\gamma_a^{d_a(o)}$.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

The identity of a transcended agent affects the utility it derives from its interactions. Also, its identity is elastic because the transcended agent can modify it over time depending on its interactions and context. If an aspect o gets a payoff of $\pi_i(o)$ in game state i, then the transcended agent derives a scaled virtual utility of $\gamma_a^{d_a(o)} * \pi_i(o)$. The overall utility of a transcended agent is computed as follows:

$$u_i(a) = \frac{1}{\sum_{\forall o \in I_a} \gamma_a^{d_a(o)}} \sum_{\forall o \in I_a} \gamma_a^{d_a(o)} \pi_i(o) \tag{2}$$

Transcended agents curate their identity by intrinsically and autonomously deciding 'whom to identify' and 'how much to identify'. They are not externally forced to form groups, collude or care about specific metrics. With multiple examples, we show that this modelling leads to emergent responsible behaviour. The limits of transcendence, factoring cost and expected utility considerations are elaborated in [7].

2.1 CT in Prisoners' Dilemma

Table 1 shows the payoff matrix of a 2-player Prisoners' Dilemma game. Irrespective of the choice of the opponent, it makes sense for both the players to defect. Thus, D is the dominant strategy and game state DD is the Nash equilibrium. Cooperation evolves only when the game is played iteratively and players have a memory of their opponent's past choices [3].

Let us introduce transcended players in this PD game. A transcended player identifies with its opponent and factors not just its own payoff but also the payoff of its opponent before making a decision. Figure 1 shows the expected utility of the two choices C and D when transcendence level, γ is varied on the $x-$axis. We observe that a low transcendence level models a selfish agent and thus $E(D) > E(C)$. However, as transcendence level increases, we note that $E(C) > E(D)$ and it is rational for the players to cooperate even in a one-shot PD game.

		Player A	
		C	D
Player B	C	R = 6, R = 6	S = 0, T = 10
	D	T = 10, S = 0	P = 1, P = 1

Table 1: Payoff Matrix for 2-player Prisoners' Dilemma

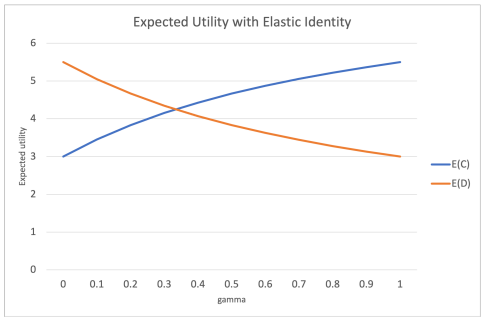

Figure 1: Expected Utility to Cooperate or Defect for a Transcended Agent in a Prisoners' Dilemma game [7]

2.2 CT in a Network

Next, we take a simple, undirected graph with nodes as agents and edges denoting possible interaction pathways between two agents. In this network, we simulate a message passing scenario where a sender sends a message to a receiver 2-hops away via an intermediate node. When the intermediate agent forwards the message, it incurs a cost and the sender gets utility as the message is received. On the other hand, when the intermediate agent drops the message, it incurs no cost whereas the sender gets a negative utility as the intended message is lost. Thus the dilemma of responsibility is faced by the intermediate agent who needs to decide whether to forward or drop a message.

In this context, we introduce transcended agents in the network who identify with their neighbourhood. They factor the utility of the sender along with their own cost and utility before deciding whether to forward or drop a message. Transcended agents in a network demonstrate responsible behaviour in the form of more messages being forwarded than dropped in the network [7]. Also, transcended agents are resilient even in the presence of adversarial agents who act selfishly.

2.3 CT versus other Paradigms of Ethics

There are a variety of paradigms of ethics like utilitarianism, deontology, virtue ethics etc. We developed a test-bed called SPECTRA on the message passing network discussed above to evaluate different paradigms of ethics including CT [5]. We demonstrate that CT as compared to other ethical paradigms, gives greater adaptability to agents such that they can demonstrate responsible behaviour to varying extents depending on the context in which they operate and their interactions with other agents in their neighbourhood.

2.4 Realistic Applications of CT

CT has also been applied to some real-world applications of multi-agent networks [6]. Supply chains can be modelled as a network of autonomous agents which decide whether to wait for more orders or dispatch existing orders. Similarly, traffic flow in a road network can be regulated using adaptive traffic lights which decide to turn green for one-of its incoming lanes and the duration of green phase. Transcended agents show promising trends in both these applications.

3 FUTURE WORK

Currently transcended agents identify with their direct neighbours in a network, we plan to extend the model so that agents can also transcend to a collection of agents or even abstract concepts like network-level metrics. So far we have done experiments in a fixed network, in future we want to try transcendence in dynamic networks like mobile adhoc networks or a network of autonomous vehicles. Applications of transcendence can be extended in multiple ways to make it more realistic and can be tested in other domains.

4 CONCLUSIONS

Computational transcendence is a framework to design autonomous agents with an elastic identity which leads to emergent responsible behaviour. We hope that CT will be useful to build responsible autonomous agents across a variety of applications.

ACKNOWLEDGMENTS

The author would like to thank Prof. Srinath Srinivasa for his constant guidance and support as advisor. The author also thanks the Machine Intelligence and Robotics (MINRO) center funded by Government of Karnataka, India and the Center for Internet of Ethical Things (CIET) funded by Government of Karnataka, India and World Economic Forum for supporting this work.

REFERENCES

[1] Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology* 7, 3 (2005), 149–155.

[2] Colin Allen, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems* 21, 4 (2006), 12–17.

[3] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science* 211, 4489 (1981), 1390–1396.

[4] Rutger Bregman. 2020. *Humankind: A hopeful history.* Bloomsbury Publishing.

[5] Janvi Chhabra, Karthik Sama, Jayati Deshmukh, and Srinath Srinivasa. 2023. Comparative Modeling of Ethical Constructs in Autonomous Decision Making. (1 2023). https://doi.org/10.36227/techrxiv.21802302.v1

[6] Jayati Deshmukh, Nikitha Adivi, and Srinath Srinivasa. 2023. Modeling Application Scenarios for Responsible Autonomy using Computational Transcendence (extended abstract, to appear). In *AAMAS*.

[7] Jayati Deshmukh and Srinath Srinivasa. 2022. Computational Transcendence: Responsibility and agency. *Frontiers in Robotics and AI* 9 (2022). https://doi.org/10.3389/frobt.2022.977303

[8] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 4698–4704. https://doi.org/10.24963/ijcai.2017/655

[9] Eugenio Oliveira, Klaus Fischer, and Olga Stepankova. 1999. Multi-agent systems: which research for which applications. *Robotics and Autonomous Systems* 27, 1-2 (1999), 91–106.

[10] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. 2021. Trustworthy human-AI partnerships. *Iscience* 24, 8 (2021), 102891.

[11] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–38.