Engineering Normative and Cognitive Agents with Emotions and Values

Doctoral Consortium

Sz-Ting Tzeng
North Carolina State University
Raleigh, United States
stzeng@ncsu.edu

ABSTRACT

While Artificial Intelligence (AI) has become part of our daily lives, there are emerging expectations for these AI systems to (1) reason over human factors with humans-in-the-loop (2) adapt to the changing environment or requirement in the real world. To achieve this goal, an agent must first incorporate emotions in its decision-making. Further, the agent must be capable of interpreting normative information from expressed emotions. Specifically, expressed emotions as information enable inference of non-observable mental states [16]. Furthermore, the agent must consider human values and preferences. The research presented here proposes an agent architecture to accommodate these factors.

KEYWORDS

Social Norms; Emotions; Reinforcement Learning; Values; Preferences

ACM Reference Format:
Sz-Ting Tzeng. 2022. Engineering Normative and Cognitive Agents with Emotions and Values: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

With advances in technology, Artificial Intelligence (AI) has become part of our daily lives. Unlike the past, the software is no longer limited to confined and isolated environments. Nowadays, software interacts with its environment, each other, and humans [4]. Therefore, as in Figure 1, humans and AI then form a multi-agent system (MAS). With humans-in-the-loop, there are emerging needs for modern AI systems to consider human factors. Specifically, these AI systems should reason over humans' behaviors. Human values help to explain behaviors and attitudes from a motivational basis [11]. Human values and preferences define an individual's intrinsic motivation and dominate how this individual thinks and evaluates everything. AI that incorporates human values would be more realistic and trustworthy.

While humans' decision-making includes internal and external attitudes, the other key factor in decision-making is emotions. Herbert Simon, a Nobel laureate, emphasized that general thinking and problem-solving must incorporate the influence of emotions [13]. Emotions, the responses to internal or external events or objects, provide extra information in communication and also serve as sanctions and social norms themselves. Social norms regulate behaviors [10, 14] in an agent society, but agents have the capability to deviate from norms in certain contexts. While introduced to multiagent systems , social norms act as societal principles or behavioral constraints that regulate agent behaviors within MAS by measuring our perceived psychological distance [1, 3, 9]. Norms either are established or revised in a top-down manner or emerge in a bottom-up manner [6, 10]. Norms from the top-down approach, such as laws, are defined by a centralized authority. Conversely, norms can also emerge from the bottom up via agent interactions. In both approaches, norms and the environment can change over time and bring out the problem of adaptability. By regulating agent interactions, norms facilitate coordination in MAS. To reduce human interventions, adaptation for AI systems becomes necessary. Including both norms and emotions helps us to build explainable and trustworthy AI.

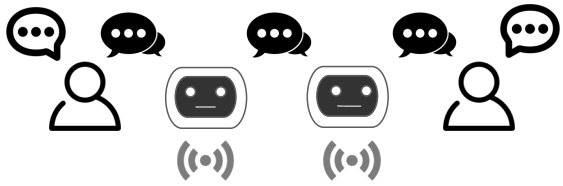

Figure 1: Real-world multiagent system

Sanctions, the reactions to norm satisfaction or norm violation, have guided research on norms for a long time. Current research on norms focuses on how sanctioning shapes agents' behaviors. Sanctions in the real world are often more subtle than mere rewards or punishments [7]. In particular, verbal messages or expressed emotions also serve as sanctions. Therefore, we investigate our first research question: How does emotional response to agent interactions affect norm emergence? We included emotions in the normative reasoning process, which evaluates and decides whether to comply or violate norms. Furthermore, explicit messages or emotional expressions can convey normative information. We then brought out our second question: How does provide indirect information, e.g., emotion as information, influence norm emergence? To address this question, we considered expressed emotions as information. To reduce human intervention and efforts, we investigate whether reinforcement learning can accommodate reasoning about cognitive constructs, emotions, and norms as our third research question. We show that reinforcement learning has the potential

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

to model norms and emotions via considering normative information as an intrinsic reward. While personal preference and values guide behaviors, we investigate how social value orientation (SVO), the preference over resource allocation between self and others, influences normative behaviors in our current work.

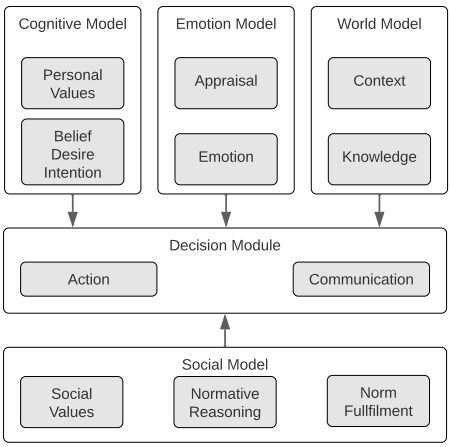

Figure 2: Agent architecture, representing and reasoning over beliefs, desires, intentions, emotions, and norms

2 AGENT ARCHITECTURE

To accommodate human factors, we propose an agent architecture as Figure 2 that consists of three components: cognitive architecture, world model, and social model.

In psychology, decision-making [12] is a cognitive process that selects a belief or a series of actions based on values, preferences, and beliefs to achieve specific goals. Our cognitive architecture describes an agent's beliefs, desires, intentions, and personal preferences. For the emotion model, we adopt the OCC model of emotions [8]. The world model describes the contexts in which an agent stands and represents the agent's general knowledge about the world. The social model of an agent includes social values, normative reasoning, and norm fulfillment. Social values define standards that individuals and groups employ to shape the form of social order, e.g., fairness and justice. The normative-reasoning component of an agent reasons over observations, norms, and possible outcomes of satisfying or violating norms. Norm fulfillment checks if a norm has been fulfilled or violated with the selected action. Sanctions may come after norm fulfillments or violations.

2.1 Expressed Emotions As Sanctions

In our first study [15], we investigate the following research question. $RQ_{emotion}$. How does modeling the emotional responses of agents to the outcomes of interactions affect norm emergence and social welfare in an agent society?

To address $\mathrm{RQ}_{\mathrm{emotion}}$, we refine the abstract normative emotional agent architecture [2] and investigate how emotions enforce norms. To make the problem tractable, we apply one social norm in our evaluation and simplify the emotional expression to reduce the complexity. Specifically, after the norm fulfillment in Figure 2, the emotion model appraises the compliance or violation of a norm and triggers emotions. We simulate a line-up scenario and observe that the triggered self-directed and other-directed emotions further enforce norms compared to agents sanctioned by predefined norms. Our findings indicate that incorporating emotions enables agents to cooperate better than those who do not.

2.2 Expressed Emotions As Information

In our second study, we investigate the following research questions.

$\mathrm{RQ}_{\mathrm{RL}}$. How does reinforcement learning accommodate reasoning about cognitive constructs, emotions, and norms?

$\mathrm{RQ}_{\mathrm{information}}$. How does providing indirect information, e.g., emotion as information, influence norm emergence?

To address these questions, we consider normative information as belief rewards and apply belief reward shaping [5], a reward augmentation framework that considers rewards from the environment and also from beliefs. We simulate a pandemic scenario and find that (1) reinforcement learning can model norms and emotions, (2) normative information from expressed emotions encourages cooperation and enforce norms.

2.3 Ongoing Work

We investigate the following research question. RQ_{SVO}. How does social value orientation influence the robustness of norms?

We include SVO into our agent framework to address this RQ. We simulate a pandemic scenario with a selfish agent society and a mixed agent society with four different kinds of SVOs: altruistic, prosocial, individualistic, and competitive. In this scenario, agents decide whether to wear masks based on their individual and SVO when interacting with other agents. The results show that agents in the mixed agent society receive higher social experience than agents in the selfish agent society. In return, agents in the mixed agent society have a higher tendency to sacrifice their preference to achieve higher collective rewards. Another concern of agents considering SVO, specifically competitive agents, may decrease the harmony of a society.

3 FUTURE WORK

While humans evaluate social norms based on their values or preferences, they accept exceptions. We hope to build agents that incorporate human values and are adaptive to changing environments and evolving norms. Further, we will investigate how do explanations shape social norms. Specifically, what information to reveal to persuade others for inevitable norm violation is what we aim to investigate, which can be referred to as the communication component in Figure 2.

ACKNOWLEDGMENT

The author would like to thank her advisor, Dr. Munindar P. Singh, and collaborator, Dr. Nirav Ajmeri, for their tremendous guidance and support. The materials presented here are based on works supported by the US National Science Foundation under grant IIS-2116751.

REFERENCES

[1] Nirav Ajmeri, Pradeep K. Murukannaiah, Hui Guo, and Munindar P. Singh. 2017. Arnor: Modeling Social Intelligence via Norms to Engineer Privacy-Aware Personal Agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, São Paulo, 230–238. https://doi.org/10.5555/3091125.3091163

[2] Estefania Argente, Elena Del Val, Daniel Perez-Garcia, and Vicente Botti. 2020. Normative Emotional Agents: A Viewpoint Paper. *IEEE Transactions on Affective Computing* (2020). https://doi.org/10.1109/TAFFC.2020.3028512

[3] Christopher D. Hollander and Annie S. Wu. 2011. The Current State of Normative Agent-Based Systems. *Journal of Artificial Societies and Social Simulation* 14, 2 (2011), 6. https://doi.org/10.18564/jasss.1750

[4] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. 2016. Revani: Revising and Verifying Normative Specifications for Privacy. *IEEE Intelligent Systems (IS)* 31, 5 (Sep 2016), 8–15. https://doi.org/10.1109/MIS.2016.89

[5] Ofir Marom and Benjamin Rosman. 2018. Belief Reward Shaping in Reinforcement Learning. In *Proceedings of the 32nd Conference on Artificial Intelligence (AAAI)*. AAAI Press, New Orleans, 3762–3769. https://doi.org/10.5555/3504035.3504496

[6] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm Emergence in Multiagent Systems: A Viewpoint paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)* 33, 6 (2019), 706–749. https://doi.org/10.1007/s10458-019-09422-0

[7] Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. 2016. Classifying Sanctions and Designing a Conceptual Sanctioning Process Model for Socio-Technical Systems. *The Knowledge Engineering Review (KER)* 31, 2 (mar 2016), 142–166. https://doi.org/10.1017/S0269888916000023

[8] Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York. https://doi.org/10.1017/CBO9780511571299

[9] Rudolph J. Rummel. 1975. Understanding Conflict and War: Vol. 1: The Dynamic Psychological Field. *Sage Publications* (1975).

[10] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7, 1 (2011), 21–54. https://doi.org/10.3233/MGS-2011-0167

[11] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919. https://doi.org/10.9707/2307-0919.1116

[12] Herbert A. Simon. 1960. *The New Science of Management Decision*. Harper & Brothers. https://doi.org/10.1037/13978-000

[13] Herbert A. Simon. 1967. Motivational and Emotional Controls of Cognition. *Psychological Review* 74, 1 (1967), 29–39. https://doi.org/10.1037/h0024127

[14] Munindar P. Singh. 2013. Norms as a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1 (dec 2013), 21:1–21:23. https://doi.org/10.1145/2542182.2542203

[15] Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. 2021. Noe: Norms Emergence and Robustness Based on Emotions in Multiagent Systems. In *Preproceedings of the International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE)*. London, 1–17. https://arxiv.org/abs/2104.15034

[16] Yang Wu, Chris L. Baker, Joshua B. Tenenbaum, and Laura E Schulz. 2018. Rational Inference of Beliefs and Desires From Emotional Expressions. *Cognitive Science* 42, 3 (2018), 850–884. https://doi.org/10.1111/cogs.12548