Engineering Pro-social Values in Autonomous Agents – Collective and Individual Perspectives

Nieves Montes(✉)

Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
nmontes@iiia.csic.es

Abstract. This doctoral thesis is concerned with the engineering of values with an explicit *pro-social* (as opposed to a personal) focus. To do so, two approaches are explored, each dealing with a different level at which interactions are studied and engineered in a multi-agent system. The first, referred to as the *collective* approach, leverages *prescriptive norms* as the promoting mechanisms of pro-social values. The second, referred to as the *individual* approach, deals with the internal reasoning scheme of agents and endows them with the ability to reason about others. This results in empathetic autonomous agents, who are able to take the perspective of a peer and understand the motivations behind their behaviour.

Keywords: Values in AI · Normative MAS · Social AI

1 Introduction

The focus of this thesis is the development of complementary approaches to engineer moral values with an explicit *pro-social focus* in autonomous agents and multi-agent systems (MAS). This includes values that seek to promote the greater good of the community, such as universalism, benevolence, and tradition. To achieve this goal, two components are explored as potential avenues to embed such pro-social values: the *prescriptive norms* that apply to a MAS as a whole, and the *individual cognitive machinery* that is triggered in direct agent-to-agent interactions. I refer to the former as the *collective approach*, and to the latter as the *individual approach*.

2 The Collective Approach

The collective approach leverages societal level constructs, in particular *prescriptive norms*, to engineer pro-social values into societies of autonomous agents. In

This work is funded by the EU TAILOR project (H2020 #952215).

D. Baumeister and J. Rothe (Eds.): EUMAS 2022, LNAI 13442, pp. 431–434, 2022.
https://doi.org/10.1007/978-3-031-20614-6_26

this line of research, we have first proposed a general methodology for the automated synthesis of prescriptive norms based on their degree of alignment with respect to some value [7,8]. There, norms are tied to optimisable parameters. This enables us to use off-the-shelf meta-heuristic techniques to find the set of norms that most successfully promote some value. Moreover, we also provide an analytic toolkit to examine the resulting optimal normative systems: the Shapley values of individual norms (which quantify the contribution of a single norm towards the alignment), and the compatibility among values (which quantifies to what degree the aggressive promotion of value v_i may hinder the achievement of a different value v_j).

Despite the progress made in [7], it has one major limitation: its rigid representation of norms requires to define the space of normative systems from scratch every time the methodology is to be used in a new scenario. To tackle this limitation, we have defined the Action Situation Language (ASL) [5,6], inspired by Elinor Ostrom's Institutional Analysis and Development framework [10].

The ASL is a logical language, implemented in Prolog, that allows communities of agents to represent a wide variety of norms in a machine-readable and syntactically-friendly way (as `if-then-where` statements). The ASL is complemented by a game engine, which takes as input a rule configuration description and automatically builds its formal semantics as an extensive-form game. This model, then, can be analysed using standard game-theoretical tools.

Overall, ASL and its complementary game engine provide a complete connection from the set of norms and regulations in place to the outcomes most incentivised by them and, consequently, the values that are being promoted by these outcomes. After this computation has been performed, the community of agents can decide whether the most likely outcomes are aligned with respect to the values most important for them. Using ASL, we have been able to model several benchmark social scenarios from the policy analysis literature. For example, we have been able to demonstrate the eradication of violent outcomes once announcement rules are introduced in a fisher community.

The ASL follows in the footsteps of previous languages for the systematic definition of extensive-form games [4,11]. However, the main feature that sets ASL apart is the fact that ASL descriptions are meant to be *extensible*. Its full power is leveraged when the effects of adding, retracting, or changing the priorities of rules (which indicate the precedence of rule statements when conflicts arise) are assessed in an automated fashion.

3 The Individual Approach

In contrast to the collective approach, the individual approach focuses on the cognitive machinery that individual agents must possess in order to abide by socially-focused values. In particular, we are interested with the values of *cooperation* and *empathy*.

To embed empathetic attitudes into autonomous agents, we are developing an agent model that combines two techniques (or families of techniques): Theory

of Mind and abductive reasoning. Theory of Mind (ToM) refers to the human cognitive ability to put oneself in the shoes of someone else and reason from their perspective. Within AI, ToM approaches are often referred to as *modelling others* and they are most prevalent in competitive domains [9]. Meanwhile, abduction refers to the logic reasoning scheme that derives, given an input observation, the best explanation for it.

The basic model consists of an observer agent i, operating with logic program T_i, and an acting agent j, operating with logic program T_j. The interaction begins when i is notified that j has selected some action a_j to perform. Then, the observer i engages in ToM by simulating the perspective that the actor j has of the state of the system at the point where they concluded that a_j was the action to perform. This means that i substitutes their program T_i by the program they estimate that j is working with, which we denote by $T_{i,j}$. In general, $T_{i,j}$ is incomplete, as i can, in general, only construct an approximation of the view that j has of the state of the system. Next, the observer i computes, using abductive reasoning, the explanations that would justify j selecting a_j. These explanations contain additional knowledge that the observer i incorporates back into their own knowledge base, to make use of them for later decision-making.

We have developed this model in Jason [2], an agent-oriented programming language. We provide a complete domain-independent implementation. Furthermore, we have tested it successfully for the cooperative card game of Hanabi. Hanabi is an award-winning card game where agents must collaborate to build stacks of cards with identical colour, however they can only see the cards of others and not their own. Players can share information with one another through hints, however doing so will spend one information token, which can later be recovered.

There are several features of Hanabi that make it an excellent benchmark to test techniques for modelling others in collaborative settings. First, Hanabi is a purely cooperative game where agents all share a common goal and need to coordinate as a team to achieve it. Second, agents have to deal with imperfect information, as they do not have access to their own cards. Therefore, there is additional information to be gained by deriving and incorporating the knowledge that peers were relying upon to select their actions. Third, information itself is collectively managed by the team as a collective resource. All of these features have led some researches to propose Hanabi as the next major challenge to be undertaken by the AI community [1], especially as interest on social and cooperative AI grows [3].

4 Conclusion

In summary, my research deals with approaches to embed socially-oriented values (i.e. those related to the greater good of the community) into autonomous agents. Two avenues are being explored to this end, which correspond to the two levels at which interactions in a multi-agent system take place: the collective level (through prescriptive norms that make up the institutional environment where

are group of agents are embedded), and the individual level (that engineers the cognitive machinery of individual agents).

Work in cooperative aspects of AI is gathering increasingly more attention, as researchers realize that AI systems are deployed in communities including other software agents and humans, and should be designed with this realization in mind [3]. The multi-agent systems community is uniquely well-positioned as this shift in focus takes place. Therefore, I believe that work seeking to embed pro-social values and mutually beneficial behaviour is highly relevant, important, and timely.

Acknowledgements. The author would like to thank her supervisors, Dr Nardine Osman and Dr Carles Sierra, for their continued guidance and support.

References

1. Bard, N., et al.: The Hanabi challenge: a new frontier for AI research. Artif. Intell. **280**, 103216 (2020). https://doi.org/10.1016/j.artint.2019.103216
2. Bordini, R.H., Hübner, J.F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason. Wiley, Hoboken (2007)
3. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: machines must learn to find common ground. Nature **593**(7857), 33–36 (2021). https://doi.org/10.1038/d41586-021-01170-0
4. Koller, D., Pfeffer, A.: Representations and solutions for game-theoretic problems. Artif. Intell. **94**(1–2), 167–215 (1997). https://doi.org/10.1016/S0004-3702(97)00023-4
5. Montes, N., Nardine, O., Sierra, C.: A computational model of Ostrom's institutional analysis and development framework. Artif. Intell. 103756 (2022). https://doi.org/10.1016/j.artint.2022.103756
6. Montes, N., Osman, N., Sierra, C.: Enabling game-theoretical analysis of social rules. In: Frontiers in Artificial Intelligence and Applications, vol. 339, pp. 90–99. IOS Press (2021). https://doi.org/10.3233/FAIA210120
7. Montes, N., Sierra, C.: Value-guided synthesis of parametric normative systems. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, pp. 907–915. AAMAS 2021, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2021). https://dl.acm.org/doi/10.5555/3463952.3464060. (Best paper award finalist)
8. Montes, N., Sierra, C.: Synthesis and properties of optimally value-aligned normative systems. J. Artif. Intell Res. (2022)
9. Nashed, S., Zilberstein, S.: A survey of opponent modeling in adversarial domains. J. Artif. Intell. Res. **73**, 277–327 (2022). https://doi.org/10.1613/jair.1.12889
10. Ostrom, E.: Understanding Institutional Diversity. Princeton University Press (2005)
11. Schiffel, S., Thielscher, M.: Representing and reasoning about the rules of general games with imperfect information. J. Artif. Intell. Res. **49**, 171–206 (2014). https://doi.org/10.1613/jair.4115