Human Influences on Decision Making in Multi-Agent Systems

Doctoral Consortium

Daniel E. Collins
University of Bristol
Bristol, United Kingdom
daniel.collins@bristol.ac.uk

ABSTRACT

For safe deployment in the real world, autonomous agents must be capable of reliably achieving and sustaining collective good while operating in dynamic environments alongside humans and other technical agents. In this paper, I detail ongoing research that explores how modelling real-world influences on human behaviour can inform the development of novel approaches for fostering pro-social collective behaviour and beneficial social outcomes in multi-agent systems.

KEYWORDS

Social Values; Affective State; Intrinsic Motivation

ACM Reference Format:
Daniel E. Collins. 2025. Human Influences on Decision Making in Multi-Agent Systems: 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

For successful real-world deployment of multi-agent systems (MAS), autonomous agents must be engineered with advanced capabilities for learning desirable social behaviours and sustaining beneficial collective outcomes while operating in sociotechnical systems [20], alongside humans and technical agents. In such environments, autonomous agents must navigate interactions between actors and groups of actors with different goals, roles and capabilities, including human stakeholders with heterogeneous social, cognitive and normative factors influencing their behaviour. These complexities pose challenges for achieving beneficial collective behaviour. Conflicting objectives among self-interested agents in mixed-motive [15] settings give rise to social dilemmas and suboptimal social outcomes. Further, real-world scenarios feature unique environmental pressures and population characteristics that shape latent incentive structures and evolve over time. To avoid unintended consequences, approaches for promoting pro-social behaviour and collective good in real-world scenarios must demonstrate robustness with respect to these influences by (1) ensuring adequate generalisation across environmental and population conditions, and (2) enabling runtime adaptation to achieve and maintain desirable social outcomes under changing conditions.

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).

Recent works have demonstrated diverse approaches for promoting pro-sociality in MAS by modelling human characteristics and social mechanisms, such as reputation [2], indirect reciprocity [21], identification with others [5], social norms [1, 17], social preferences [15], and ethics [22]. For reinforcement learning (RL) agents, intrinsic motivation approaches support the learning of adaptive behaviours in the absence of hand-crafted rules or rewards, reflecting innate human drives [8, 10, 18]. While there is a rich body of research exploring the influence of cognitive factors such as beliefs, knowledge, and emotional reasoning on human behaviour [6], less attention has been given to modelling their interplay alongside social influences and environmental pressures in MAS.

In this paper, I present my doctoral research exploring two complementary approaches for promoting beneficial collective behaviour in MAS: (1) modelling the socially adaptive interplay between dynamic emotions and static social preferences in multi-agent decision making, and (2) a framework for promoting cooperation through intrinsically motivated responsibility. I conclude by discussing these works in the context of broader themes around behavioural influences and incentives, and future research directions towards developing robust pro-social MAS.

2 SOCIAL VALUE ORIENTATION AND INTEGRAL EMOTION

Social Value Orientation (SVO) [14] is a spectrum of measurable personality traits that characterise preferences for individual versus collective welfare. Individual differences between characteristics such as SVO offer insight into the heterogeneity of human behaviour in the real world. However, individual differences alone cannot explain the dynamic nature of human decision making, e.g., adapting preferences in response to changing circumstances. Human decision making is better characterised by the interplay between stable individual differences and dynamic, contextual influences [6]. This is demonstrated in research exploring the influence of integral emotions (IE) on human decision making — task-related affective states that arise directly from observations in the current decision-making context [13]. Once evoked, IEs strongly influence behaviour, often overriding the other social and cognitive influences. In [3], we investigated whether modelling IEs in populations of agents with heterogeneous baseline SVO policies could help to achieve more equitable collective outcomes in multi-agent decision making. We hypothesised that by enabling agents to deviate from their baseline policy based on experience, the suboptimality and negative social impact of certain SVOs across different interactions can be diluted, resulting in improved social outcomes across populations with different distributions of SVO. To test this, we developed

Svoie, a method for adjusting the probability that agents act according to their baseline SVO policy versus alternative IE policies depending on their current emotional "valence", an internal state representing the positiveness or negativeness of the integral emotion resulting from recent task outcomes. While SVOs are used to define baseline heterogeneous social preferences, *Svoie* enables temporary deviations from social preferences depending on valence. We evaluated *Svoie* through simulation experiments in Colored Trails (CT) [7], an iterated negotiation game designed to study social decision making. We model IE valence as an internal state that evolves based on the outcomes of repeated rounds of CT with random opponents sampled from populations with heterogeneous SVO. Poor task performance increases the probability of spiteful behaviour [24], modelled as maximizing relative advantage over opponents. Conversely, high performance increases the probability of adopting a passive, inequity-averse strategy. We evaluated *Svoie* in simulated populations with different distributions of SVO. We found that *Svoie* populations consistently achieved a statistically significant reduction in welfare inequality compared to populations of baseline SVO agents. These findings suggest that methods enabling run-time behavioural adaptation based on task outcomes and other contextual influences could support improved collective outcomes in heterogeneous agent populations, presenting a promising direction for future work.

3 IMPLICIT RESPONSIBILITY

In [4], we introduce "implicit responsibility" (IR), a novel form of intrinsically motivated responsibility, capturing a key aspect of emergent cooperation in the real world — the natural tendency to assume responsibility for others' welfare in certain scenarios, without explicit external incentives. Specifically, we consider a form of IR that describes a realistic intrinsically-motivated pro-social behaviour — the tendency to act on opportunities to help others when the personal cost and risk of doing so is minimal. We outline conditions necessary for the emergence of this form of pro-social IR in a conceptual framework, and argue that learning to recognise these conditions independently may support emergent cooperation, complementing existing responsibility approaches based on explicit norms and commitments [23]. Based on this framework, we developed and evaluated a novel reward-shaping approach for fostering cooperation among self-interested reinforcement learning (RL) agents. In this implementation, we operationalise pro-social IR conditions as environment-specific rules for the formation and violation of IRs. We model IR agents as RL agents that self-penalise for violating IRs — cases where the agent could have helped another agent but failed to do so — via reward-shaping, analogous to models of guilt or regret. We evaluated IR agents against RL baselines through experiments in a mixed-motive environment, where agents must forage for resources to survive. We designed this environment so that, under certain parametrisations, mutual cooperation via strategic resource-sharing provides greater individual and collective welfare than purely self-interested or altruistic strategies. We found that, in this setting, our IR agents learned optimal cooperative strategies with greater sample efficiency compared to baseline RL agents. These findings suggest that modelling implicit social responsibilities can accelerate the emergence of cooperation among self-interested agents. However, further experimentation and analysis is needed to validate these findings and better characterise our approach. In planned work, we aim to repeat our experiments across different parametrisations of our environment, capturing different underlying environmental incentives for resource sharing versus independent foraging.

4 FUTURE DIRECTIONS

The high-level aim of my research is to explore novel avenues for designing and evaluating methods that promote pro-social behaviour and collective good in MAS. These contributions highlight three key directions for future work: (1) modelling the interplay between normative, social and cognitive influences on decision making in MAS; (2) developing methods for fostering pro-social behaviours and sustaining collective good across varied scenarios with different environmental pressures and population characteristics; (3) developing approaches that enable agents to adapt to real-time changes in environmental and social dynamics.

In [4], environment-specific rules define conditions for forming and violating IRs, limiting application to different environmental conditions and scenarios. In ongoing work, I am investigating the use of "empowerment" [12, 19] — a measurement based on information-theoretic formalism that can serve as a proxy for quantifying an agents' potential influence over future states of the world — as the basis of IR reasoning, to remove the dependence on environment-specific rules, and improve generalisation.

In work currently under review, we present a design framework for creating customisable multi-agent environments from reusable components. Specifically, we include design features that enable systematic variation of parameters that control the environment setup and logic. We aim to use this framework to evaluate the robustness of methods for incentivising pro-social behaviour by assessing their efficacy, generalisability and side effects under variations in environment dynamics and population characteristics.

In [3], we touch upon an important capability for real-world decision making — adapting behaviour online according to experience. Future work will explore techniques from multi-objective RL (MARL) [9] that support online adaptation to dynamic environment conditions. For example, [11, 16] demonstrate an approach for training multi-objective policies that can be conditioned at run-time to approximate single-objective policies, i.e., different weighted combinations of multiple reward functions. Exploring such procedures in the context of objectives that capture human influences and motivations presents an intriguing avenue for future work, e.g., for characterising the alignment between behaviours learned via scenario-specific extrinsic reward functions and information-theoretic objectives across scenarios with different underlying incentive structures.

ACKNOWLEDGMENTS

DC was supported by the UK Research and Innovation (UKRI) Centre for Doctoral Training in Interactive Artificial Intelligence Award (EP/S022937/1). DC thanks their supervisors, Dr Nirav Ajmeri and Dr Conor J Houghton, for their invaluable advice and support.

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 16–24. https://doi.org/10.5555/3398761.3398769

[2] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 115–123.

[3] Daniel E. Collins, Conor J. Houghton, and Nirav Ajmeri. 2023. Social Value Orientation and Integral Emotions in Multi-Agent Systems. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE) (Lecture Notes in Computer Science)*. Springer, London, 118–138. https://doi.org/10.1007/978-3-031-49133-7_7

[4] Daniel E. Collins, Conor J. Houghton, and Nirav Ajmeri. 2024. Fostering Multi-Agent Cooperation through Implicit Responsibility. In *Proceedings of the 2nd International Workshop on Citizen-Centric Multiagent Systems (CMAS)*. Auckland, 1–10.

[5] Jayati Deshmukh. 2023. Emergent Responsible Autonomy in Multi-Agent Systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 3029–3031.

[6] Alia El Bolock. 2020. *What Is Character Computing?* Springer International Publishing, Cham, 1–16. https://doi.org/10.1007/978-3-030-15954-2_1

[7] Sevan G. Ficici, Avi Pfeffer, Ya'akov Gal, Barbara Grosz, and Stuart Shieber. 2008. Colored trails: a multiagent system testbed for decision-making research. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Demo Papers* (Estoril, Portugal) *(AAMAS '08)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1661–1662.

[8] Christian Guckelsberger, Christoph Salge, and Simon Colton. 2016. Intrinsically motivated general companion NPCs via Coupled Empowerment Maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)* (Santorini, Greece). IEEE Press, 1–8. https://doi.org/10.1109/CIG.2016.7860406

[9] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Kallstrom, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. 2023. A Brief Guide to Multi-Objective Reinforcement Learning and Planning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1988–1990.

[10] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019 (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 3040–3049. http://proceedings.mlr.press/v97/jaques19a.html

[11] Johan Källström and Fredrik Heintz. 2019. Tunable Dynamics in Agent-Based Simulation using Multi-Objective Reinforcement Learning. In *Proceedings of the 2019 Adaptive and Learning Agents Workshop (ALA), 2019* :. 1–7.

[12] Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. 2005. All Else Being Equal Be Empowered. In *Advances in Artificial Life*, Mathieu S. Capcarrère, Alex A. Freitas, Peter J. Bentley, Colin G. Johnson, and Jon Timmis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 744–753.

[13] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and Decision Making. *Annual Review of Psychology* 66, 1 (Jan. 2015), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

[14] Charles G. McClintock and Scott T. Allison. 1989. Social Value Orientation and Helping Behavior. *Journal of Applied Social Psychology* 19, 4 (1989), 353–362.

[15] Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duèñez Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 869–877.

[16] David O'Callaghan and Patrick Mannion. 2021. Tunable Behaviours in Sequential Social Dilemmas using Multi-Objective Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1610–1612.

[17] Ninell Oldenburg and Tan Zhi-Xuan. 2024. Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1510–1520.

[18] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 2778–2787.

[19] Christoph Salge, Cornelius Glackin, and Daniel Polani. 2014. *Empowerment–An Introduction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 67–114. https://doi.org/10.1007/978-3-642-53734-9_4

[20] Munindar P. Singh. 2013. Norms As a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1, Article 21 (Dec. 2013), 23 pages.

[21] Martin Smit and Fernando P. Santos. 2024. Learning Fair Cooperation in Mixed-Motive Games with Indirect Reciprocity. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence (IJCAI-2024)*. International Joint Conferences on Artificial Intelligence Organization, 220–228. https://doi.org/10.24963/ijcai.2024/25

[22] Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. 2025. Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, Philadelphia, 1–9.

[23] Vahid Yazdanpanah, Enrico Gerding, Sebastian Stein, Corina Cirstea, M.C. Schraefel, Timothy James Norman, and Nick Jennings. 2021. Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements. *IEEE Internet Computing* 6, 25 (2021), 15–22. https://doi.org/10.1109/MIC.2021.3107334

[24] Ya Zheng, Zhong Yang, Chunlan Jin, Yue Qi, and Xun Liu. 2017. The Influence of Emotion on Fairness-Related Decision Making: A Critical Review of Theories and Evidence. *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.01592