Ethical Decision-Making in Multi-Agent Systems

Jessica Woodgate
University of Bristol
Bristol, United Kingdom
jessica.woodgate@bristol.ac.uk

ABSTRACT

Developing socially beneficial multi-agent systems (MAS) necessitates addressing the capacity of agents to make decisions of an ethical nature. Ethics is inherently multi-agent, involving one's concern for another. To make ethical decisions, agents should consider the needs of different stakeholders. Principles from normative ethics, the philosophical study of morality, provide practical guidance to determine right from wrong. My work implements normative ethics principles in artificial agents to foster ethical decision-making.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**; *Cooperation and coordination.*

KEYWORDS

normative ethics; social norms; sociotechnical systems; fairness

ACM Reference Format:
Jessica Woodgate. 2025. Ethical Decision-Making in Multi-Agent Systems. 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

Multi-agent systems (MAS) are collections of multiple agents acting and interacting in a shared environment [39]. Decisions in MAS with numerous interacting agents have varying effects on outcomes for relevant stakeholders. An agent may negatively impact others if it prioritises solely its own interests and does not consider others. To be beneficial to the system, an agent should consider relevant stakeholders in its decisions. Given that ethics is multi-agent insofar as it involves one party's concern for another [29], it is important to consider the capacity of agents in MAS to make decisions of an ethical nature [12].

In my work, I argue that operationalising principles from normative ethics, the philosophical study of practical means to determine right from wrong [4], is a step towards developing agents with ethical decision-making capacities. Normative ethics principles denote and justify morally good or right action [20]. Principles guide decision-makers in making evaluative judgements and determining moral permissibility of courses of actions, facilitating choosing amongst actions by considering their moral implications [23, 26].

Normative ethics principles are justified in philosophical theory; normative in the sense that they are prescriptive, denoting how things ought to be, rather than descriptive, denoting how things are. As what is the case might not be ethical, using independently justified principles has the benefit of addressing the *is-ought* gap: just because something *is* the case, doesn't mean it *ought* to be [21]. Implementing normative ethics principles makes explicit the reasons underlying ethical choices, as to explain why a decision was made, one can refer to the reasons that justify the relevant principle [12, 17]. Operationalising normative ethics thus provides a mechanism to systematically assess the rightness and wrongness of actions in a range of situations, and justify decisions by reference to the principles used [3].

In pursuit of operationalising normative ethics for ethical decision-making in MAS, I identify three key research questions to address:
RQ$_{\text{identification}}$ *What ethical principles have previously been implemented in computer science literature?* Understanding how to apply principles, and which principle is appropriate to a particular situation, is aided by identifying how computer science literature has previously utilised normative ethics. To address this question, we surveyed computer science literature and developed a taxonomy of 21 principles operationalised in AI [42].
RQ$_{\text{operationalisation}}$ *How can ethical principles be operationalised in decision-making capacities?* Operationalising ethical principles in decision-making assists in choosing amongst possible actions [7, 23]. We addressed this question by implementing maximin (a well known fairness principle prioritising the least advantaged [32]) [44], and combining multiple principles in learning agents [43].
RQ$_{\text{incorporating context}}$ *How does context interplay with the application of ethical principles to MAS?* Particular settings may have requirements that influence the relevancy of various factors and outcomes of decisions. Continuing work pursues questions regarding the influence of context on ethical decision-making.

2 PRINCIPLE IDENTIFICATION

In normative ethics, there are many theories about morality with varying strengths and weaknesses. All good theories have some useful truths, yet different principles can lead to distinct solutions in the same situation, and all principles have some counter-intuitive implications [12, 34]. To support ethical reasoning in the face of imperfect principles, a reasonable response is to use each principle where it is most effective [5]. Discerning which principle is appropriate for an application is aided by identifying the principles that have previously been used in literature and how they have been implemented. In Woodgate and Ajmeri [42], we survey computer science literature and develop a taxonomy of 21 normative ethical principles previously operationalised. We define a new mapping of each principle to how it has been operationalised, key themes practitioners should be aware of to implement principles, difficulties that may arise, and existing gaps.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

3 PRINCIPLE OPERATIONALISATION

Implementing principles in reproducible ways requires consistent methodologies that make explicit the principles being used [10]. To investigate how to implement normative ethics in decision-making capacities in MAS, we (1) operationalise Rawlsian ethics to foster fair norm emergence; (2) combine multiple principles to reconcile difficulties arising with individual principles.

3.1 Operationalising Rawlsian Ethics

Social norms are standards of expected behaviour [27]. Norms have been harnessed in MAS to regulate behaviour. However, exploitative norms may emerge when agents act solely out of self-interest. In Woodgate et al. [44] we present RAWL·E, a novel method to design socially intelligent norm-learning agents that consider others in decision-making by operationalising maximin, a fairness principle advanced by Rawls [32]. Maximin states that in a society with unequal distribution not to the benefit of all, the least well-off should be prioritised. Previous literature utilises principles to aggregate value preferences [22, 36], make normative decisions [2], and optimise learning policies [13, 38]. We advance previous literature by applying maximin to learning agents in norm emergence settings. We find societies of RAWL·E agents have higher fairness and social welfare, and more emerged cooperative norms, compared to societies of agents not implementing maximin.

3.2 Operationalising Multiple Principles

Implementing multiple principles in decision-making helps to see problems from different perspectives [26], and balance the strengths and weaknesses of each principle [6]. Principles have been implemented in MAS, however, prior work does not combine principles, combines principles in a single way, or presumes a central authority, which may not be feasible in all environments [9, 13, 25, 30, 45]. To mitigate weaknesses with individual principles, in Woodgate and Ajmeri [43] we propose PriENE, an agent architecture combining multiple principles in individual decision-making. We evaluate a society of PriENE agents and societies implementing individual principles in a berry harvesting scenario. We find societies of PriENE agents have higher fairness and sustainability than societies implementing single principles [43].

4 INCORPORATING CONTEXT

Challenges arise with implementing ethical decision-making in the real world, as there are varying factors that affect outcomes of decisions and how to apply principles. There are multiple ways to choose between or combine principles, people may reasonably disagree about morality, decisions are made within historical contexts with distinct power dynamics, and decisions should be interpretable to stakeholders. Planned work and future directions investigate the interplay between context and ethical decision-making.

4.1 Planned Work

Deciding which principles to encode in decision-making is a challenging task: there are issues with individual principles, and different principles may have conflicting recommendations [31]. In Woodgate and Ajmeri [43], we investigated how combining principles into a single answer mitigates weaknesses with individual principles. We found different ways of combining principles may be appropriate for distinct scenarios. Directions include examining the influence of context on which principles or combination of principles is appropriate. Directions also involve combining promotion of ethical behaviour with explicit prevention of unethical outcomes.

Even if it were possible to identify one principle that held true in any situation, humans hold a variety of reasonable and contrasting beliefs [33]. Rational people may disagree about descriptive facts (the mechanics of a situation), preferences (agree about descriptive facts but want different things), or what is of moral value (what is right or wrong) [34]. Designing AI with one moral doctrine may therefore impose beliefs upon people who do not agree with them [18]. Directions involve investigating how ethical decision-making can take into account beliefs and preferences of stakeholders.

To evaluate the methods developed in Woodgate et al. [44] and Woodgate and Ajmeri [43], we simulated abstracted berry harvesting scenarios. To improve real world applicability, directions include applying methods to more complex and real-world scenarios.

4.2 Future Directions

Ethical MAS should promote fairness, broadly understood as the mitigation of bias and discrimination against marginalised groups [19]. There has been extensive research into algorithmic fairness, however, focusing on algorithms alone can subvert actual fairness by taking too narrow a stance [17]. Developing tools that support fairness and are socially beneficial should prioritise the experience of the people that are affected by those tools. Future directions include exploring participatory approaches under a sociotechnical lens that appreciates interacting social and technical tiers [29, 41].

People often have differing and potentially conflicting preferences, and multiple-user social dilemmas may arise when values (deeply held beliefs and preferences [35]) or norms conflict [11, 40]. Dilemmas may arise in mundane settings, and do not have to be extreme trolley-problem cases [15]. Previous research has explored dilemmas where one principal makes a decision affecting others (one-to-many) [16]. However, there are several ways in which power dynamics affect how agents act and interact. Interactions could involve one agent affecting another (one-to-one), many agents affecting one (many-to-one), or many agents affecting many others (many-to-many). Future directions include investigating various types of multiple-user social dilemmas, including the role of social commitments and notion of spheres of commitment [8, 37].

In social contexts, decisions should be interpretable insofar as stakeholders can infer some sort of qualitative understanding [14]. Interpretability is important as the reasons for a decision help evaluate that decision [24, 28]. Qualitative understanding may be aided through explanations, illocutionary acts uttered with the intention to make something understandable [1]. Agents should have the ability to justify their decisions so that stakeholders can understand why those decisions were made. Future directions include examining how ethical decision-making in STS can be made interpretable, and how explanations can be harnessed to improve interpretability.

ACKNOWLEDGMENTS

Thanks to Nirav Ajmeri and Paul Marshall for the support, and EPSRC Doctoral Training Partnership Grant No. EP/W524414/1.

REFERENCES

[1] Peter Achinstein. 1977. What Is an Explanation? *American Philosophical Quarterly* 14, 1 (1977), 1–15. http://www.jstor.org/stable/20009644

[2] 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

[3] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAccT)*, Vol. 81. PMLR, New York, 149–159.

[4] Paula Boddington. 2023. *Normative Ethical Theory and AI Ethics.* Springer Nature Singapore, Singapore, 229–276. https://doi.org/10.1007/978-981-19-9382-4_6

[5] Jason Brennan. 2007. *The best moral theory ever: The merits and methodology of moral theorizing.* Ph.D. Dissertation. University of Arizona.

[6] Emanuelle Burton, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, and Toby Walsh. 2017. Ethical Considerations in Artificial Intelligence Courses. *AI Magazine* 38, 2 (July 2017), 22–34. https://doi.org/10.1609/aimag.v38i2.2731

[7] Cansu Canca. 2020. Operationalizing AI Ethics Principles. *Commun. ACM* 63, 12 (Nov. 2020), 18–21. https://doi.org/10.1145/3430368

[8] Cristiano Castelfranchi. 1995. Commitments: From Individual Intentions to Groups and Organizations.. In *Proceedings of the 1st International Conference on Multiagent Systems*. AAAI, San Fransico, 41–48.

[9] Janvi Chhabra, Karthik Sama, Jayati Deshmukh, and Srinath Srinivasa. 2024. Evaluating computational models of ethics for autonomous decision making. *AI and Ethics* 1 (2024), 1–14. Issue 1. https://doi.org/10.1007/s43681-024-00532-4

[10] Vincent Conitzer, Walter Sinnott-Armstrong, J. S. Borg, Yuan Deng, and Max Kramer. 2017. Moral Decision Making Frameworks for Artificial Intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, Honolulu, 4831–4835.

[11] Virginia Dignum. 2019. *Ethical Decision-Making.* Springer, Cham, 35–46. https://doi.org/10.1007/978-3-030-30371-6_3

[12] Virginia Dignum. 2019. *Responsible Artificial Intelligence.* Number 1 in Artificial Intelligence: Foundations, Theory, and Algorithms. Springer Cham, New York. https://doi.org/10.1007/978-3-030-30371-6

[13] Shaokang Dong, Chao Li, Shangdong Yang, Bo An, Wenbin Li, and Yang Gao. 2024. Egoism, utilitarianism and egalitarianism in multi-agent reinforcement learning. *Neural Networks* 178 (2024), 106544. https://doi.org/10.1016/j.neunet.2024.106544

[14] Adrian Erasmus, Tyler D. P Brunet, and Eyal Fisher. 2021. What is Interpretability? *Philosophy & technology* 34 (2021), 833–862. Issue 4. https://doi.org/10.1007/s13347-020-00435-2

[15] Amitai Etzioni and Oren Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21 (Dec. 2017), 403–418. Issue 4. https://doi.org/10.1007/s10892-017-9252-2

[16] Anja K Faulhaber, Anke Dittmer, Felix Blind, Maximilian A Wächter, Silja Timm, Leon Sütfeld, Sütfeld LR, Achim Stephan, Gordon Pipa, and Peter König. 2019. Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and engineering ethics* 1, 1 (April 2019), 399–418. https://doi.org/10.1007/s11948-018-0020-x

[17] Sina Fazelpour, Zachary C. Lipton, and David Danks. 2022. Algorithmic Fairness and the Situated Dynamics of Justice. *Canadian Journal of Philosophy* 52, 1 (2022), 44–60. https://doi.org/10.1017/can.2021.24

[18] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (Sept. 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[19] Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter L. Spirtes, Yang Liu, and Lu Cheng. 2024. Long-Term Fairness Inquiries and Pursuits in Machine Learning: A Survey of Notions, Methods, and Challenges. arXiv:2406.06736 [cs.LG] https://arxiv.org/abs/2406.06736

[20] Svantje Guinebert. 2020. How do moral theories stand to each other? *Zeitschrift für Ethik und Moralphilosophie* 3, 2 (Oct. 2020), 279–299. https://doi.org/10.1007/s42048-020-00077-1

[21] Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *JAIR* 70 (May 2021), 871–890. https://doi.org/10.1613/jair.1.12481

[22] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite López-Sánchez, and Juan Rodríguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through lp-Regression. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, New Zealand, 780–788.

[23] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. 2019. Moral Permissibility of Action Plans. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)* 33, 01 (July 2019), 7635–7642. https://doi.org/10.1609/aaai.v33i01.33017635

[24] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, Stockholm, 3122–3130. https://proceedings.mlr.press/v80/lipton18a.html

[25] Mehdi Mashayekhi, Nirav Ajmeri, George F. List, and Munindar P. Singh. 2022. Prosocial Norm Emergence in Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 17, 1–2 (June 2022), 3:1–3:24. https://doi.org/10.1145/3540202

[26] Bruce M. McLaren. 2003. Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence* 150, 1 (2003), 145–181. https://doi.org/10.1016/S0004-3702(03)00135-8 AI and Law.

[27] 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.

[28] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116

[29] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 1706–1710. https://doi.org/10.5555/3398761.3398958 Blue Sky Ideas Track.

[30] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically Compliant Planning within Moral Communities. In *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Virtual Event, 188–198. https://doi.org/10.1145/3461702.3462522

[31] Erik Persson and Maria Hedlund. 2022. The future of AI in our hands? To what extent are we as individuals morally responsible for guiding the development of AI in a desirable direction? *AI and Ethics* 2, 4 (Nov. 2022), 683–695. https://doi.org/10.1007/s43681-021-00125-5

[32] John Rawls. 1958. Justice as Fairness. *The Philosophical Review* 67, 2 (April 1958), 164–194. https://doi.org/10.2307/2182612

[33] John Rawls. 1999. *A Theory of Justice: Revised Edition.* Harvard University Press, Harvard.

[34] Pamela Robinson. 2023. Moral disagreement and artificial intelligence. *AI and Society* 38, 3 (June 2023), 1–14. https://doi.org/10.1007/s00146-023-01697-y

[35] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[36] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Stefano Moretti. 2023. Value alignment in participatory budgeting. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, London, 1692–1700. https://doi.org/10.5555/3635637.3663030

[37] Munindar P. Singh. 1999. An ontology for commitments in multiagent systems. *Artificial Intelligence and Law* 7, 1 (1999), 97–113. https://doi.org/10.1023/A:1008319631231

[38] Justin Svegliato, Samer Nashed, and Shlomo Zilberstein. 2020. An Integrated Approach to Moral Autonomous Systems. *Frontiers in Artificial Intelligence and Applications* 325 (2020), 2941 – 2942. https://doi.org/10.3233/FAIA200464

[39] Gerhard Weiss. 2013. *Multiagent systems* (second edition. ed.). The MIT Press, Cambridge, Massachusetts ;. http://site.ebrary.com/id/10674446

[40] Jessica Woodgate. 2023. Ethical Principles for Reasoning about Value Preferences. In *Proceedings of the 6th AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, Montréal, 972–974. https://doi.org/10.1145/3600211.3604728

[41] Jessica Woodgate and Nirav Ajmeri. 2022. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Online, 1824–1828. https://doi.org/10.5555/3535850.3536118 Blue Sky Ideas Track.

[42] Jessica Woodgate and Nirav Ajmeri. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. *CSUR* 56, 289 (July 2024), 1–37. https://doi.org/10.1145/3672394

[43] Jessica Woodgate and Nirav Ajmeri. 2025. Combining Normative Ethics Principles to Learn Prosocial Behaviour. In *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Detroit, 3.

[44] Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. 2025. Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, Philadelphia, 1–8.

[45] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, Online, 12967–12978.