Ethical Principles for Reasoning about Value Preferences

Jessica Woodgate
University of Bristol
Bristol, United Kingdom
yp19484@bristol.ac.uk

ABSTRACT

To ensure alignment with human interests, AI must consider the preferences of stakeholders, which includes reasoning about values and norms. However, stakeholders may have different preferences, and dilemmas can arise concerning conflicting values or norms. My work applies normative ethical principles to resolve dilemma scenarios in satisfactory ways that promote fairness.

KEYWORDS

normative ethical principles, values, norms, sociotechnical systems, fairness

ACM Reference Format:
Jessica Woodgate. 2023. Ethical Principles for Reasoning about Value Preferences. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 8–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3600211.3604728

1 INTRODUCTION

Multiagent systems (MAS) understood as sociotechnical systems (STS) consist of multiple human-agent duos, with a social tier that imposes regulations upon a technical tier [20, 21]. To improve fairness considerations, it is important to appreciate the interaction of multiple users, rather than single agents [6]. When viewing STS from this holistic perspective, stakeholders govern by promoting norms that align with their values. However, issues arise when stakeholders have different preferences, or where values or norms conflict [10]. Decisions must be made that consider stakeholder preferences, values, and norms in ways that promote fairness.

Previous work examines using values to reason about norms. Kayal et al. [12] develop a normative conflict resolution model based on value profiles of users, which selects norms that best support the stakeholders' values. Montes and Sierra [19] provide a methodology for evaluating the value alignment of norms by examining changing preferences. However, often not all stakeholders will agree on which factor is the most important in a given scenario [9]. In these dilemmas, there may be cases where multiple norms conflict with each other, one or more norms conflict with the value preferences of a user, or value preferences of one user conflict with those of other users. There may also be scenarios in which values and norms do not conflict, however a decision must be made that fairly considers a variety of different preferences.

AIES '23, August 8–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604728

Challenges thus remain in creating decision support for everyday dilemmas in which there are differing preferences, or values or norms conflict, aiming towards the development of systems with fair governance. The application of ethical principles may improve fairness considerations in aggregating value preferences.

Resolving these dilemmas in satisfactory ways with a higher goal of fairness may be achieved by operationalising normative ethical principles in decision support [25]. Normative ethics is the study of practical means to determine the ethicality of an action [7, 21]. Leben [14] provides foundations for mechanising certain ethical principles, which could applied to decision support in STS. Normative ethical principles have also been operationalised in domains such as resource allocation and machine ethics [5, 9, 14, 23].

2 RESEARCH QUESTIONS

RQ_1 **What ethical principles currently exist in computer science literature?** A framework operationalising principles may help to methodically analyse scenarios and promote satisfactory outcomes [9]. A taxonomy identifying and categorising ethical principles in computer science literature would aid the development of this framework. This taxonomy could then be expanded to principles seen in philosophy and other disciplines.
RQ_2 **How can ethical principles be operationalised in reasoning capacities needed to govern STS?** Developing methods to incorporate ethical principles in reasoning techniques used to govern STS would be beneficial to support ethical decision making.
RQ_3 **How can context be incorporated in the application of ethical principles?** Ethical decision making is context dependent, and which principles are appropriate to apply in specific circumstances may vary. Methods to incorporate context could improve the applicability of principles.

3 COMPLETED WORK: TAXONOMY OF NORMATIVE ETHICAL PRINCIPLES FOR AI

To address RQ_1, we have developed a taxonomy of normative ethical principles previously used in computer science literature.
Motivation. Ethical principles can support decisions as they help to guide normative analysis, understand different perspectives, and determine the moral permissibility of concrete courses of actions [15, 18, 23]. A framework aiding the operationalisation of principles in decision making may be useful to methodically think through scenarios and promote satisfactory outcomes [9]. To create such a framework, it is first necessary to identify and categorise ethical principles previously seen in computer science literature.
Background. Related work includes Tolmeijer et al. [24] which studies how principles relate to machine ethics, and Yu et al. [27] which proposes a taxonomy of ethical decision frameworks. As ethical thinking should be fostered through appreciating various

approaches [4], expanding these works to incorporate a wider variety of principles, and how they have been operationalised, may improve the amplitude of ethical reasoning. A larger taxonomy of principles that currently exist in computer science literature, examining how each principle has been operationalised, could help form the groundwork for an ethical decision support framework.

Completed Work. Following the guidelines of Kitchenham et al. [13], we conducted a systematic literature review of computer science literature. We developed a taxonomy of 23 normative ethical principles operationalised in AI [26]. We describe how each principle has previously been operationalised, highlighting key themes AI practitioners seeking to implement ethical principles should be aware of. Future directions involve looking outside of the domain of ethics used in computer science, to examine ethical theories in philosophy and other disciplines. This includes researching principles from cultures outside of the Western doctrine, which may aid better application to groups of stakeholders from diverse backgrounds.

Contribution. Broadening the range of ethical principles found in previous surveys, we identify a taxonomy tree with 23 ethical principles discussed in Computer Science literature. Principle specific operationalisation is presented, with new mapping of each principle to how they have been operationalised in literature [26].

4 ONGOING WORK: OPERATIONALISING ETHICAL PRINCIPLES

To address RQ_2, we are developing a model that operationalises normative ethical principles in reasoning about value preferences.

Motivation. When values are imbued in systems, aggregating values into a single outcome may improve ethical decision making in STS [22]. However, reasoning about values is challenging [17], and stakeholders could have personal preferences between different values [16, 21]. Value preferences of some stakeholders may conflict with value preferences of other stakeholders, or values may conflict with norms [10, 25].

Background. Previous work integrates normative ethics in decision making, and utilises values to reason about norms. Cointe et al. [7] propose an agent which utilises normative ethical theories to improve ethical decision making in MAS, which could be expanded to consider value preferences of multiple stakeholders. Montes and Sierra [19] provide a methodology for examining the alignment of norms with values. To expand this, the application of ethical principles may improve fairness considerations in aggregating values to help resolve conflicts. Ajmeri et al. [2] aggregate value preferences of users, applying a single normative ethical principle. However, to resolve scenarios in which principles lead to unintuitive outcomes, or are unable to promote one action over another, it is important to apply a variety of different principles.

Current Work. Our current work lays the foundations for a model demonstrating how multiple ethical principles can be implemented in reasoning about values of stakeholders. In our model, agents have value preferences for the payoffs they receive. Different ethical principles are applied to these value preferences to reach a decision which promotes fairness. Via an example of smart heating STS scenario, we demonstrate how we could apply our model. Each stakeholder has an internal hierarchy of individual value preferences [19]. At each timestep, all agents propose their preferences, and a collective decision is made by applying different ethical principles to those preferences. We conduct preliminary simulation experiments on our model. To evaluate the emergence of norms that promote fairness, we compute quality metrics in each run of the simulation including health, wealth, and Gini coefficient.

Preliminary Results. Preliminary results suggest the most appropriate ethical principle to apply in a situation may depend on the metrics being used, as different principles can lead to different outcomes. We find that the principle best suited to maximise payoffs is the principle of Maximin. However, if a fair distribution of resources is more important, the most appropriate principle is Egalitarianism. These findings may help the development of agents that can learn the best principle to apply in certain situations.

Contribution. Incorporating ethical principles in reasoning, considering the preferences of stakeholders. This may improve fairness considerations in aggregating different value preferences and resolving value conflicts. Applying multiple ethical principles may help to view dilemmas from different perspectives and improve the amplitude of ethical reasoning.

5 NEXT STEPS: INCORPORATING CONTEXT

To address RQ_3, there are several directions future work could address to improve the contextual applicability of principles.

- **Considering Contextual Value Preferences.** Our current work assumes each stakeholder's order of value preferences is fixed. However, preferences may change [8, 17]. Future work could involve expanding our current simulations to incorporate contextual values and changing value preferences.
- **Incorporating Internal Reasoning in Agents.** In our current work agents do not have internal reasoning schemes, as decisions are deferred to a collective decision making module. Future work could include equipping agents with internal reasoning, so that aggregating individual ethical decision making using normative ethical principles can be studied on an individual level.
- **Resolving Conflicts Between Ethical Principles.** Our preliminary results suggest that different principles might be appropriate in different scenarios. Sometimes a single ethical principle may lead to an unintuitive outcome, or be unable to give a clear preference between two different options. When seeking the best principle to apply, it is important that agents can consider several different principles to identify a suitable solution [7, 26]. Future work includes developing learning agents that can resolve conflicts between different principles and optimise the application of principles. By incorporating explainability in these agents, we can investigate how agents learn to handle such scenarios [1].
- **Using Logic to Encode Ethical Principles** Our current work applies abstracted versions of ethical principles to demonstrate the basic idea of how such principles may be applied to reason about value preferences. To achieve more precise representations and improve contextual applicability, future work could utilise logic techniques such as those used by Govindarajulu and Bringsjord [11] and Berreby et al. [3] to encode normative ethical principles in the governance of STS.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council Doctoral Training Partnership [EP/W524414/1].

REFERENCES

[1] Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. 2022. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Vienna, 10–14.

[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] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2017. A Declarative Modular Framework for Representing and Applying Ethical Principles. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, São Paulo, 96–104. https://doi.org/10.5555/3091125.3091145

[4] 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

[5] Violet (Xinying) Chen and J. N. Hooker. 2020. A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism. In *Proceedings of the 3rd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, New York, 221–227. https://doi.org/10.1145/3375627.3375844

[6] Amit Chopra and Munindar Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 1st AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, New Orleans, 48–53. https://doi.org/10.1145/3278721.3278740

[7] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, Singapore, 1106–1114.

[8] Daniel Collins, Conor Houghton, and Nirav Ajmeri. 2023. Social Value Orientation and Integral Emotions in Multi-Agent Systems. arXiv:2305.05549 [cs.MA]

[9] 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.

[10] Virginia Dignum and Frank Dignum. 2020. Agents Are Dead. Long Live Agents!. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, New Zealand, 1701–1705. https://doi.org/10.5555/3398761.3398957

[11] Naveen Sundar Govindarajulu and Selmer Bringsjord. 2017. On Automating the Doctrine of Double Effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, Melbourne, 4722–4730. https://doi.org/10.24963/ijcai.2017/658

[12] Alex Kayal, Willem-Paul Brinkman, Mark A. Neerincx, and M. Birna van Riemsdijk. 2018. Automatic Resolution of Normative Conflicts in Supportive Technology based on user values. *ACM Transactions on Internet Technology (TOIT)* 18, 4, Article 41 (May 2018), 21 pages.

[13] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University and Durham University Joint Report. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[14] Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, New York, 86–92. https://doi.org/10.1145/3375627.3375808

[15] 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

[16] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Pradeep K. Murukannaiah. 2023. Value Inference in Sociotechnical Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, London, 1774–1780.

[17] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, London, 799–808. https://doi.org/10.5555/3463952.3464048

[18] 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.

[19] Nieves Montes and Carles Sierra. 2021. Value-Guided Synthesis of Parametric Normative Systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Virtual Event, London, 907–915.

[20] 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.

[21] Pradeep K. Murukannaiah and Munindar P. Singh. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE Internet Computing* 24, 3 (May 2020), 51–57. https://doi.org/10.1109/MIC.2020.2989935

[22] Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. 2022. Design Heuristics for Ethical Online Institutions. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV*, Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu (Eds.). Springer International Publishing, Cham, 213–230.

[23] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating Ethics within Machine Learning Courses. *ACM Transactions on Computing Education* 19, 4 (Aug. 2019), 1–26. https://doi.org/10.1145/3341164

[24] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2021. Implementations in Machine Ethics: A Survey. *CSUR* 53, 6, Article 132 (Dec. 2021), 38 pages. https://doi.org/10.1145/3419633

[25] 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.

[26] Jessica Woodgate and Nirav Ajmeri. 2022. Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. *arXiv* 2208.12616 (Aug. 2022), 1–37. arXiv:2208.12616 [cs.CY] https://arxiv.org/abs/2208.12616

[27] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, Stockholm, 5527–5533. https://doi.org/10.24963/ijcai.2018/779