Unlocking the Potential of Machine Ethics with Explainability

Extended Abstract

Timo Speith
University of Bayreuth
Bayreuth, Germany
timo.speith@uni-bayreuth.de

ABSTRACT

Roughly speaking, the research field of machine ethics deals with devising behavioral constraints on computational systems to ensure restricted, morally acceptable behavior. The potential benefits of researching machine ethics are substantial, encompassing contributions to ethical AI development and the societal impact of computational systems. However, there are genuine concerns and risks associated with this research (e.g., the potential to undermine human autonomy) that must be carefully considered.

In this article, we will explore the question of whether it is worthwhile to conduct research in machine ethics, given the potential demerits and challenges involved. Central to our study is the proposition that explainability, such as is being explored in connection with explainable artificial intelligence (XAI), can serve as a powerful tool to augment the advantages of machine ethics research, mitigate its disadvantages, and create unique advantages of its own. Overall, we conclude that the study of machine ethics is worthwhile, especially when it is supported by research on explainability.

KEYWORDS

Machine Ethics; Ethical AI; Explainability; Explainable AI; XAI

ACM Reference Format:
Timo Speith. 2024. Unlocking the Potential of Machine Ethics with Explainability: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Artificial systems permeate our world. A vital question arises from this permeation: How should we constrain machines to behave in a morally acceptable way towards us humans? This question concerns *machine ethics*—the search for formal, unambiguous, algorithmizable, and implementable behavioral constraints on systems so as to compel them to exhibit morally acceptable behavior [6, 11, 38, 39].

Unfortunately, there seem to be potential demerits connected with research in machine ethics. For instance, some people believe that as artificial systems are equipped with increasingly sophisticated capacities for moral reasoning, they inevitably acquire a moral status at some point [17, 20]. Such a moral status would imply that we have to grant rights to machines, resulting in extremely negative consequences for a society that relies on exploiting them.

In this extended abstract (for the full paper, see Chapter 3 of [38]), we will outline that the reasons for pursuing the research discipline of machine ethics outweigh the reasons against doing so. Afterwards, we will argue that explainability can serve as a catalyst for machine ethics, profoundly contributing to it.

2 IS MACHINE ETHICS WORTHWHILE?

2.1 Motivations for Machine Ethics

Motivation 1: The Well-Being of Humankind. To ensure the well-being of people and prevent them from being harmed by artificial systems, it is essential to develop systems that act in accordance with morality [4, 17, 33].

Motivation 1.1: Acceptance. Giving systems the capacity to reason morally is often assumed to be one prerequisite for making them trustworthy and, thus, increasing their acceptance [6, 26].

Motivation 1.2: Fairness. Machine ethics can help to prevent implicit biases from being programmed into a system, since implementing morals presumably requires making values explicit [6, 7].

Motivation 1.3: Benign Superintelligence. One option to increase the chances of a benign superintelligence emerging is to equip artificial systems with capacity for moral reasoning [24, 33].

Motivation 2: Moral Alignment. Engaging in machine ethics may improve the moral alignment of decision makers beyond current standards [3, 4, 17]. We may become able to formulate better moral theories and, perhaps, even find the correct one (if it exists).

Motivation 2.1: Improved Human Decisions. Studying machine ethics can help to improve individual human decisions [3, 6]. Humans may use a machine to arrive at moral solutions just as they use pocket calculators to arrive at mathematical solutions [17, 39].

Motivation 2.2: Improved Human Morality. Machine ethics can help to improve human morality as a whole. Engaging in machine ethics can help us to formulate more consistent moral theories and reach consensus on moral dilemmas [4, 6–9, 17, 35].

2.2 Risks of Machine Ethics

Risk 1: Insufficiently Moral Machines. As ethics was originally devised for humans, we need to investigate whether it is possible and makes sense to apply results from ethics to machines.

Risk 1.1: Lacking Moral Agency. In contrast to humans, machines are not moral agents and, thus, cannot act morally. Machines not being moral agents is the case because machines lack relevant capacities such as free will, consciousness, or autonomy [3, 4, 33].

Risk 1.2: Unacceptable Level of Morality. For machines to be accepted, it is plausible that a higher level of morality is required than that for humans [3, 4, 17]. This level might be unattainable [17].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Risk 1.3: Computational Limitations. It may be impossible to implement an adequate moral theory in a machine due to theoretical and practical computational limitations [3, 4, 16, 17, 35].

Risk 2: Detrimental Consequences. Equipping machines with the capacity to reason morally may indirectly (Risks 2.1.1 and 2.1.2) or directly (Risks 2.2.1–2.2.4) lead to undesirable consequences.

Risk 2.1.1: Increased Corruptibility. Moral capacities may require an explicit representation of moral considerations that can make the machine more susceptible to errors or corruption [17, 40].

Risk 2.1.2: Moral Patiency. Endowing systems with moral capacities may lead to them becoming moral patients [17, 20]. Such systems would have certain rights that we must not violate [23, 25].

Risk 2.2.1: Bad Moral Performance. With added moral capacities, the overall performance of systems may suffer [4] while still not reaching human standards of moral performance [4, 16, 17].

Risk 2.2.2: Responsibility Gap. System deployment goes hand-in-hand with problems of attributing responsibility [4, 13, 17, 29]. Deploying machines with moral capacities could make attributing responsibility even more difficult than before [15, 17, 19, 37].

Risk 2.2.3: Value Imperialism. Equipping systems with a particular capacity to reason morally can violate national or cultural identity, leading to some kind of value imperialism [17].

Risk 2.2.4: Undermined Human Agency. Equipping machines with moral capacities may undermine human agency by correcting their (moral) mistakes and thus supporting human incompetence [17].

2.3 Refuting the Risks of Machine Ethics

To refute the reasons against machine ethics, we champion *moral alignment* [17]. According to this concept, the goal of machine ethics is to make the behavior of machines more morally desirable from the perspective of humans, even if only by a little bit [17, 35, 36].

Risk 1: Insufficiently Moral Machines. With moral alignment, we do not need to care about lacking moral agency, an unacceptable level of morality, or unimplementable moral theories. Pragmatically seen, more and more machines will be deployed, and a slightly morally desirable machine deployed is morally better than one that is not morally desirable but deployed nevertheless.

Risk 2: Detrimental Consequences. Many of the risks mentioned above arise only when one tries to give machines capacities for being moral in the strong sense humans are. With moral alignment, however, we do not need to give machines such capacities.

3 THE ADVANTAGES OF EXPLAINABILITY

While machine ethics alone is worthy of exploration, machine explainability can provide further support. In a nutshell, machine explainability is concerned with making various aspects of an artificial system understandable to a stakeholder [10, 18, 27, 34].

3.1 Amending the Risks of Machine Ethics

Machine explainability can help to further mitigate the risks of machine ethics. We focus on the risks for which this applies most.

Risk 2.1.1: Increased Corruptibility. Explanations of a system's inner workings can help pinpoint sources of errors or corruption and, thus, enable developers to improve and fix them [14, 32].

Risk 2.2.1: Bad Moral Performance. Where unacceptable outcomes occur, explanations can help identify where a machine's moral capacities are defective and need to be adjusted.

Risk 2.2.2: Responsibility Gap. One of the central motivations for pursuing machine explainability is to be better able to attribute responsibility [13, 32, 34]. For an in-depth discussion, see [13].

Risk 2.2.4: Undermined Human Agency. Explanations let humans (regain) control over a situation. As a consequence, humans remain responsible in this situation and uphold their agency [32].

3.2 Augmenting the Motivations of Machine Ethics

Machine explainability can significantly augment each of the reasons for pursuing the research discipline of machine ethics.

Motivation 1: The Well-Being of Humankind. Calibrating the acceptance of machines [14, 22] as well as fairness [1, 2, 10] are central motivations for machine explainability [18, 27].

Motivation 2: Moral Alignment. By explaining their solutions, artificial systems might improve individual human decisions or they might educate humans about morality as a whole.

3.3 New Advantages for Machine Ethics

Machine explainability does not only avert risks of machine ethics and augments its advantages, but is also beneficial on its own.

Advantage 1: Machine Ethics Acceptance. In addition to the acceptance of *systems*, machine explainability can promote the acceptance of *machine ethics* itself [11, 12, 38].

Advantage 2: Improved Machine Morality. Machine explainability can help us to improve systems (see Section 3.1). As the capacities for moral reasoning are part of the systems, machine explainability can also help us to improve these capacities.

Advantage 3: Enriched Machine Ethics. Machine explainability can be seen as a part of machine ethics itself [17, 41], as giving explanations can be an ethical requirement [41].

4 CONCLUSION AND FUTURE WORK

Machine ethics and machine explainability are deeply intertwined, as machine ethics needs machine explainability to unlock its full potential. While we have not discussed the other direction, machine explainability can also benefit from machine ethics. For example, the moral constraints that a morally aligned system plausibly contains could serve as the basis for the generated explanations (see [38]).

Future research needs to show how machine ethics and machine explainability can be usefully combined to achieve the best synergistic effects. One possibility in this direction could be argument-based decision making (see, e.g., [5, 11, 12, 28, 38]). The advantage of such an approach to machine decision making would be that it mimics one way in which humans come to decisions [21, 30, 31].

ACKNOWLEDGMENTS

Work on this paper was funded by the Volkswagen Foundation grants AZ 9B830, AZ 98509, and AZ 98514 “Explainable Intelligent Systems” (EIS) and by the DFG grant 389792660 as part of TRR 248.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 36th Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI 2018)*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). Association for Computing Machinery, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Colin Allen, Gary Varner, and Jason Zinser. 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 3 (2000), 251–261. https://doi.org/10.1080/09528130050111428

[4] Colin Allen, Wendell Wallach, and Iva Smit. 2006. Why Machine Ethics? *IEEE Intelligent Systems* 21, 4 (2006), 12–17. https://doi.org/10.1109/MIS.2006.83

[5] Leila Amgoud and Henri Prade. 2009. Using Arguments for Making and Explaining Decisions. *Artificial Intelligence* 173, 3–4 (2009), 413–436. https://doi.org/10.1016/j.artint.2008.11.006

[6] Michael Anderson, Susan Leigh Anderson, and Chris Armen. 2004. Towards Machine Ethics. In *Proceedings of the AAAI Workshop on Agent Organizations: Theory and Practice* (San Jose, California, USA) *(AAAI WS 2004)*, Virginia Dignum, Daniel Corkill, Catholijn Jonker, and Frank Dignum (Eds.). AAAI Press, Palo Alto, CA, USA, 53–59. https://aaai.org/Library/Workshops/2004/ws04-02-008.php

[7] Michael Anderson, Susan Leigh Anderson, and Chris Armen. 2006. An Approach to Computing Ethics. *IEEE Intelligent Systems* 21, 4 (2006), 56–63. https://doi.org/10.1109/MIS.2006.64

[8] Susan Leigh Anderson. 2008. Asimov's "Three Laws of Robotics" and Machine Metaethics. *AI & Society* 22 (2008), 477–493. https://doi.org/10.1007/s00146-007-0094-5

[9] Susan Leigh Anderson. 2011. How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better. In *Machine Ethics*. Cambridge University Press, Cambridge, England, UK, Chapter 30, 524–530. https://doi.org/10.1017/cbo9780511978036.036

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Benjamins Richard, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[11] Kevin Baum, Holger Hermanns, and Timo Speith. 2018. From Machine Ethics to Machine Explainability and Back. In *International Symposium on Artificial Intelligence and Mathematics*, Martin Charles, Dimitrios I. Diochnos, Jürgen Dix, Frederick Hoffman, and Guillermo R. Simari (Eds.). International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, USA, 1–8. https://isaim2018.cs.ou.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf

[12] Kevin Baum, Holger Hermanns, and Timo Speith. 2018. Towards a Framework Combining Machine Ethics and Machine Explainability. In *Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology*, Bernd Finkbeiner and Samantha Kleinberg (Eds.). Electronic Proceedings in Theoretical Computer Science, Sydney, NSW, AU, 34–49. https://doi.org/10.4204/EPTCS.286.4

[13] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology* 35, 1 (2022), 1–30. https://doi.org/10.1007/s13347-022-00510-w

[14] Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence* (Melbourne, Victoria, Australia) *(IJCAI XAI 2017)*, David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone (Eds.). 8–13. http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

[15] Nick Bostrom and Eliezer Yudkowsky. 2014. The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence*, Keith Frankish and William M. Ramsey (Eds.). Cambridge University Press, Cambridge, England, UK, 316–334. https://doi.org/10.1017/CBO9781139046855.020

[16] Miles Brundage. 2014. Limitations and Risks of Machine Ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26, 3 (2014), 355–372. https://doi.org/10.1080/0952813X.2014.895108

[17] Stephen Cave, Rune Nyrup, Karina Vold, and Adrian Weller. 2019. Motivations and Risks of Machine Ethics. *Proc. IEEE* 107, 3 (2019), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

[18] Larissa Chazette, Wasja Brunotte, and Timo Speith. 2021. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *Proceedings of the 29th IEEE International Requirements Engineering Conference*. IEEE, Piscataway, NJ, USA, 197–208. https://doi.org/10.1109/RE51729.2021.00025

[19] John Danaher. 2019. The Rise of the Robots and the Crisis of Moral Patiency. *AI & Society* 34, 1 (2019), 129–136. https://doi.org/10.1007/s00146-017-0773-9

[20] David Davenport. 2014. Moral Mechanisms. *Philosophy & Technology* 27, 1 (2014), 47–60. https://doi.org/10.1007/s13347-013-0147-2

[21] Irene-Anna N. Diakidoy, Loizos Michael, and Antonis Kakas. 2017. Knowledge Activation in Story Comprehension. *Journal of Cognitive Science* 18, 4 (2017), 439–471.

[22] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. Transparency You Can Trust: Transparency Requirements for Artificial Intelligence Between Legal Norms and Contextual Concerns. *Big Data & Society* 6, 1 (2019), 1–14. https://doi.org/10.1177/2053951719860542

[23] Christopher Grau. 2006. There Is No "I" in "Robot": Robots and Utilitarianism. *IEEE Intelligent Systems* 21, 4 (2006), 52–55. https://doi.org/10.1109/MIS.2006.81

[24] John S. Hall. 2011. Ethics for Self-Improving Machines. In *Machine Ethics*, Michael Anderson and Susan Leigh Anderson (Eds.). Cambridge University Press, New York, NY, USA, 512–523.

[25] Kenneth Einar Himma. 2009. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics and Information Technology* 11, 1 (2009), 19–29. https://doi.org/10.1007/s10676-008-9167-5

[26] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. 2021. On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops* (Notre Dame, Indiana, USA) *(REW 2021)*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, Piscataway, NJ, USA, 169–175. https://doi.org/10.1109/REW53955.2021.00031

[27] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296, Article 103473 (2021), 24 pages. https://doi.org/10.1016/j.artint.2021.103473

[28] Luca Longo and Lucy Hederman. 2013. Argumentation Theory for Decision Support in Health-Care: A Comparison With Machine Learning. In *Proceedings of the International Conference on Brain and Health Informatics* (Maebashi, Japan) *(BHI 2013)*, Kazuyuki Imamura, Shiro Usui, Tomoaki Shirao, Takuji Kasamatsu, Lars Schwabe, and Ning Zhong (Eds.). Springer, Berlin/Heidelberg, Germany, 168–180. https://doi.org/10.1007/978-3-319-02753-1_17

[29] Andreas Matthias. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6, 3 (2004), 175–183. https://doi.org/10.1007/s10676-004-3422-1

[30] Hugo Mercier. 2016. The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences* 20, 9 (2016), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

[31] Hugo Mercier and Dan Sperber. 2011. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences* 34 (2011), 57–111. https://doi.org/10.1017/S0140525X10000968

[32] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3, 2 (2016), 1–21. https://doi.org/10.1177/2053951716679679

[33] James H. Moor. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21, 4 (2006), 18–21. https://doi.org/10.1109/MIS.2006.80

[34] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29, 3 (2019), 441–459. https://doi.org/10.1007/s11023-019-09502-w

[35] Thomas M. Powers. 2006. Prospects for a Kantian Machine. *IEEE Intelligent Systems* 21, 4 (2006), 46–51. https://doi.org/10.1109/MIS.2006.77

[36] Thomas M. Powers. 2011. Incremental Machine Ethics. *IEEE Robotics & Automation Magazine* 18, 1 (2011), 51–58. https://doi.org/10.1109/MRA.2010.940152

[37] Daniel B. Shank, Alyssa DeSanti, and Timothy Maninger. 2019. When Are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions After Individual and Joint Decisions. *Information, Communication & Society* 22, 5 (2019), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

[38] Timo Speith. 2023. *Building bridges for better machines: from machine ethics to machine explainability and back.* Ph.D. Dissertation. Saarland University. https://doi.org/10.22028/D291-40450

[39] Steve Torrance. 2005. A Robust View of Machine Ethics. In *Proceedings of the AAAI Fall Symposium on Machine Ethics* (Arlington, Virginia, USA) *(AAAI FS 2005)*, Michael Anderson, Susan Leigh Anderson, and Chris Armen (Eds.). AAAI Press, Palo Alto, CA, USA, 88–93. https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf

[40] Dieter Vanderelst and Alan Winfield. 2018. The Dark Side of Ethical Robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, Louisiana, USA) *(AIES 2018)*, Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi (Eds.). Association for Computing Machinery, New York, NY, USA, 317–322. https://doi.org/10.1145/3278721.3278726

[41] Kate Vredenburgh. 2022. The Right to Explanation. *Journal of Political Philosophy* 30, 2 (2022), 209–229. https://doi.org/10.1111/jopp.12262