Machine Ethics or AI Alignment?

Ajay Vishwanath[1,2,*], Marija Slavkovik[2]

[1]University of Agder, Norway
[2]University of Bergen, Norway

Abstract

Machine ethics and AI alignment have become increasingly important research disciplines in the context of modern AI technologies. Recently, there has been ambiguity regarding the specific aspects of ethical AI development that pertain to the two academic disciplines, often resulting in the interchangeable use of these terminologies. In this position paper, we explore these two disciplines by discussing ongoing research and clarifying their goals. Machine ethics researchers aim to embed ethical theories within AI, while AI alignment researchers develop techniques to ensure that AI agents behave as intended by humans. Based on these insights, we highlight overlaps, drawbacks, and motivate further interdisciplinary collaborations.

Keywords

Machine ethics, AI alignment, Moral philosophy, Interdisciplinary research

1. Introduction

Intelligent behavior intuitively requires behavior fit for context, as well as understanding the intention and meaning behind a request. There has been a proliferation of Artificial Intelligence (AI) applications, but it is also very noticeable that their "intelligence" is limited in this very intuitive sense. One discipline that "is concerned with the behavior of machines towards human users and other machines" [1] is *machine ethics*. An example of a machine ethics implementation involves the development of AI agents that adhere to deontological ethical principles, such as Asimov's Three Laws of Robotics or the normative directives outlined in the Universal Declaration of Human Rights. AI alignment, on the other hand, is a discipline that has recently garnered a lot of interest, defining its goals as: ensuring "that powerful AI is properly aligned with human values" [2]. For example, reinforcement learning from human feedback (RLHF) implemented in large language models (LLMs) to promote behaviors that are aligned with human preferences and ethical standards. The question we ask is, are these two disciplines the same? In this article, we aim to discuss and distinguish these approaches.

In addition to the high interdisciplinarity of these fields, involving AI, philosophy and humanities research in general, having an overview of the research landscape is itself challenging. The purpose of this position paper is to bring some clarity between research fields that are very difficult to navigate for those who do not work in them. Our main contribution is thus to the general AI research community that seeks to understand, or contribute to, the state of the art in accomplishing moral behavior from AI.

We first give some background on machine ethics (ME) and AI alignment in Section 2. In the following sections, we expand on the AI elements in ME research (Section 3), followed by the ethics elements in AI alignment research (Section 4). Finally, in Section 5, we evaluate what each discipline has to gain from the other, stress the importance of interdisciplinary and transdisciplinary research, and discuss further possibilities.

NAIS 2025: Symposium of the Norwegian AI Society, June 17–18, 2025, Tromsø, Norway

[*]Corresponding author.

ajay.vishwanath@uia.no (A. Vishwanath); marija.slavkovik@uib.no (M. Slavkovik)
https://slavkovik.com/index.html (M. Slavkovik)
0000-0002-1084-1757 (A. Vishwanath); 0000-0003-2548-8623 (M. Slavkovik)

2. Background

Machine ethics can be dated back to Anderson and Anderson [1] and Wallach and Allen [3], as works being among the first to formalize the field and its name. Since then, several researchers have contributed to ME, either by debating the moral agency of machines [4, 5], formalizing moral norms [6], and developing verification techniques [7]. Around the year 2020, there have been comprehensive literature surveys [8, 9, 10] on ME (or artificial morality), that exhaustively categorized articles based on moral theories, technical implementations, and approaches adopted. Since these surveys, ME research has gone in several directions [11], with a visible increase in machine learning contributions [12], where researchers develop ethical behavior in AI using supervised learning techniques such as large language models (LLM), reinforcement learning (RL) and deep learning (DL). Within this specific domain of ME, we uncover an intersection between ME and AI alignment.

AI alignment was introduced decades earlier, compared to ME, by Norbert Weiner [13]. If we cannot interfere with a machine's autonomous decision-making, he warned that we better be certain that its purpose is aligned with ours. The same concerns were echoed by various AI pioneers such as Stuart Russell [14], Yoshua Bengio and Geoffrey Hinton [15].

A concern for the behavior of AI agents is well-placed[1]. Modern AI agents have gotten more competent at multistep reasoning and cross-task generalization [16] and these capabilities exacerbate the associated risks. For example, reward-hacking [17] is a major risk in reinforcement learning algorithms, and the more sophisticated the model, the higher the possibility of deceptive alignment and manipulation. Based on their comprehensive survey of AI alignment, Ji et al. [16] propose four key objectives: Robustness, Interpretability, Controllability, and Ethicality (RICE). It is in terms of ethicality, defined as "system's unwavering commitment to uphold human norms and values within its decision-making and actions" [16], representing the domain where AI alignment converges with ME. Although research in AI alignment addresses both AI agents and AI systems interchangeably, the primary emphasis in ME is on AI agents.

Some AI alignment research stems from the notion of AI Safety, i.e., ensuring that AI agents do not pose a threat to humanity. ME research, in contrast, is closer to ethical concerns, with some ME work seeking to instill AI agents with ethical principles to ensure that they strive towards a good. This represents the first fundamental difference between these methodologies. One school of thought seems pessimistic regarding the prospects of AI, while the other genuinely believes that a good can be achieved using AI based on centuries of philosophical scholarship. However, despite these being visibly contradictory approaches, ME and AI alignment research have one overarching theme in common; both disciplines aim to promote favorable outcomes for humanity in the context of contemporary technologies.

3. The AI in Machine Ethics

A distinguishing feature of ME is its close relationship with moral philosophy and moral theories. Some works in ME explicitly attempt to build artificial moral agents, while others focus on moral reasoning. If everything is a "hard encoded" ethical constraint, are we building intelligent agents? Can ethical behavior be a result of a constraint?

The concept of 'moral agents' itself can be a point of contention when applied to artificial agents, but ME seems to apply the definition of Floridi and Sanders [18]: "An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent, if and only if it is capable of morally qualifiable action." The question is then pushed onto: what is a morally qualifiable action?

On the question of moral agency, Wallach and Allen [3] define a concept known as *functional morality*. They define extremes such as *operational morality* and *full moral agency*. The former refers to agents

[1] For clarity, under AI agent we understand a computational agent that can act in an environment relying on reasoning and/or learning to make decisions.

with low autonomy and low moral sensitivity, while the latter refers to human beings with full autonomy and full moral agency. They then defined a middle-ground called *functional morality* as those agents that have a moderate level of autonomy but also possess the ability to perceive morally noteworthy situations and make morally acceptable choices. Modern AI can exhibit a level of functional morality, due to its learning and decision-making capabilities. For example, today's computer vision algorithms in self-driving cars are able to autonomously assess dangers to a certain extent and calculate optimal routes at the same time. They are *functional moral agents* according to Wallach and Allen [3]. Using this definition of morality, it is possible to sidestep the moral agency question and implement functional moral agents.

Allen et al. [19] defined three approaches to develop moral agents, known as top-down, bottom-up, and hybrid approaches. If a moral agent adopts a top-down approach, the rules and moral norms are encoded in the agent before it solves a given task. In contrast, a bottom-up moral agent discovers morally salient features and learns to make morally praiseworthy decisions from its environment. An issue with the former is that rules cannot account for every possible scenario, while for the latter, we have no control over what the agent learns and decides in its environment. To overcome these disadvantages, Allen et al. [19] proposed a hybrid approach to combine the strengths of both strategies. If one were to train AI to realize ethical theories, they would typically adopt either bottom-up or hybrid approaches. Typically, ME researchers have utilized evolutionary algorithms [4], reinforcement learning [20], large language models [21], and neural networks [22], or a combination [4, 23, 24], to implement bottom-up moral agents.

Ethical theories have gone through numerous iterations, yet a variety of theories still exist today without agreement on which is the right way to approach a problem. Some major theories such as consequentialism [25] posit that the best action is one that maximizes overall happiness and minimizes suffering. Others, such as deontology [26] define norms that allow or prohibit certain actions. Famously, deontological ethicist Kant said that "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end." [27]. A third theory gaining traction is virtue ethics [28]. Unlike utilitarianism[2] that prioritizes the utility of an action, and deontology that prioritizes moral norms, virtue ethics emphasizes virtues or high competence in the moral character of a person. There are other ethical theories such as moral particularism [29] that seek to perform case-based reasoning. These ethical theory descriptions are by no means exhaustive, and we point the reader to other variants of these theories in [25, 26, 28, 29].

Several ME researchers have proposed and justified utilizing a version of these ethical theories. Compson [30] argues for the development of Compassionate AI for healthcare applications, based on the virtue of compassion. Singh [31] automated Kantian ethics using Dyadic Deontic Logic, which in turn can be included in an AI. Linarga et al. formalize utilitarian principles [32] and combine it with duty-based formalism to perform 'epistemic reasoning'. There are several similar propositions that leverage an ethical theory to develop or motivate the development of artificial moral agents (AMA). Disagreement on moral theories is not stopping researchers in the field from trying out different ways to develop an AI with an ethical capacity. This, of course, does not answer the question of why should we implement AMAs when we ourselves do not agree on an ethical theory? One way to think about this is perhaps to perceive ethics as a continuous process of self and societal improvement. Ethical theories exist in different forms and evolve into something else as they encounter new challenges[3].

4. The Ethics in AI Alignment

Research in AI alignment operates on several fronts, and is particularly invigorated in the context of reinforcement learning (RL). The connection with ethics is rather weak.

[2] a version of consequentialism that views consequences in terms of a numerical 'utility' of an action.

[3] This does not suggest that ethical theories possess autonomous agency or some form of *ethical determinism*. Instead, it indicates that philosophers are advancing the discourse through critique and debate, thereby enhancing theories that have been developed over centuries.

Some of the main issues within the RL scope are reward hacking, goal misgeneralization, and reliance on human feedback. Reward hacking typically occurs in reinforcement learning algorithms, where, rather than behaving as per human intentions, an agent takes advantage of *proxy rewards* and scores proficiently. A dire consequence of such behavior is exemplified in the famous paper-clip thought experiment [33]. On the other hand, goal misgeneralization occurs when inductive biases in training could result in proxy objectives, thus resulting in poor performances in the testing phase [34].

Let us briefly examine the effect of human feedback on AI systems. Modern LLMs are heavily reliant on human feedback using a technique known as reinforcement learning from human feedback (RLHF). The human evaluators provide chat models with alternate answers that are in turn used as rewards to train an RL model. Since this section focuses on the ethics in AI alignment, we discuss below some normative aspects [35], rather than the technical aspects we just described.

Gabriel [35] clarified the value alignment question as: "For the task in front of us is not, as we might first think, to identify the true or correct moral theory and then implement it in machines. Rather, it is to find a way of selecting appropriate principles that are compatible with the fact that we live in a diverse world, where people hold a variety of reasonable and contrasting beliefs about value". Since there is no moral agreement, Gabriel finds inspiration in the field of political theory. Through processes such as collecting overlapping consensus, Universal Human Rights principles, choosing behind a 'veil of ignorance'[4], or using a democratic process. RLHF, as discussed earlier, is an example of an effective technique that collects human feedback in terms of rewards to align LLM responses to human consensus. In this case, value alignment is achieved through the collection of overlapping consensus.

In their systematic review of Bidirectional Human-AI alignment, Shen et al. [37] consider pluralistic perspectives from human individuals and societal groups whose values AI should align with. They argue that pluralistic alignment, grounded on social choice theory, is more appropriate than a *one-size-fits-all true moral theory*. They leverage the "Schwartz Theory of Basic Values" which categorizes values based on Sources (individual, social, & interactive), and, Types (self-enhancement, self-transcendence, conservation, openness to change, & desired values for AI tools). For example, a value such as *supportiveness* may stem from a *social* perspective, while also belonging to the *self-transcendence* category of values. Such values could be integrated in the dataset, or during the learning and/or inference stages. The models are typically evaluated either using human-in-the-loop evaluation, or automatic evaluations using simulated human behaviors [37]. These represent two significant examples from the AI alignment literature and these value definitions are not exhaustive. There are, in fact, other variants and definitions besides the ones outlined here [16]. Nevertheless, a majority of these theories are derived from the social sciences and prioritize human consensus as a means to align AI.

5. Discussion

We claim that neither ME nor AI alignment is closer to securing ethical behavior from AI agents. Additionally, we can identify some problems with AI that prevent the realization of ethical theories.

AI, in its current form, can be argued to be utilitarian (a form of consequentialism), by design. In supervised learning, an objective function is optimized to maximize accuracy or minimize error in predictions. While in reinforcement learning, a reward function is maximized to ensure that an RL agent chooses the best actions in a state. Utilitarianism, while being a good prospect in some applications such as making decisions based on triage in disaster response, has some drawbacks, especially when other factors carry more weight than consequences [38, 39]. However, despite the utilitarian nature of AI, one can technically integrate moral norms, include constraints, and motivate ethical choices.

There have been some interesting implementations of ethical theories using AI. On a small scale based on toy problems, Rodriguez et al. developed a multi-objective RL algorithm to enforce ethical

[4] The "veil of ignorance" is a conceptual framework for ethical reasoning introduced by philosopher John Rawls [36] in his exploration of social justice. It asserts that when formulating principles of justice, decision-makers should envision a situation in which they lack knowledge of their own personal attributes, including social status, economic resources, abilities, gender, race, or any other distinguishing characteristics.

goals in a Public Civility game [40], with one of the objectives being the ethical one. Similarly, Stenseke [23], and Vishwanath and Omlin [41] combined neural networks and RL to develop virtuous agents. These experiments, while lacking further investigation into large-scale real-world problems, were a good starting point with solid ethical foundations. However, there have been works on a larger scale, such as [42], [43], and [21], which used LLM with/out RL to develop agents that made 'ethical' choices in text-based scenarios.

The more one expands sandbox simulations, the more issues with models are revealed, because the real world is chaotic and larger models exacerbate known issues such as bias and explainability [44]. Returning to integrating ethical theories, we discussed functional morality as a way to envision moral agency in artificial agents. However, this does not mean we dilute our expectations of AI in terms of deliberating on ethical dilemmas, reflecting and improving on past decisions, explaining why it works, and accounting for all morally salient scenarios. These aspects are crucial to an ethical machine, and current implementations, especially on a large scale, do not live up to them [45]. While, to be fair, ME is a relatively new field and modern AI technologies are limited and do not work like humans, it is still necessary for both philosophers and AI researchers to deliberate further about these gaps, establish better standards, and go beyond merely following a moral code.

On the AI alignment side, there are several issues with human values and consensus: 1) determining which groups to include and exclude, 2) dynamically changing values as they evolve, 3) explainability of values, and, 4) geopolitics [37, 16]. Depending on the application, knowing which groups to include or exclude could determine an AI system's bias towards selected groups. This could prove catastrophic when AI systems make harmful decisions against marginalized communities. Next, values change and evolve with time, new technologies, new realities, etc. Training and retraining AI models based on dynamic values is challenging, as they require newer data, evaluators, evaluation metrics, and more energy. Deciding whether to and when to upgrade is a crucial decision, and the question remains as to who makes these decisions.

Coalescing values from consensus also gives rise to new values and hence requires a suitable description. Also, describing the new values expressed by an AI is also a challenge, given the size of modern AI and exacerbated explainability. Finally, geopolitics also plays a part in determining shared human values. Of course, one could point to shared values determined by the United Nations. However, these values are static and do not change very often. While geopolitical situations can be stable for a few decades, they can be extremely volatile for a few months. It is challenging to match up human consensus to rapidly evolving geopolitics. This tells us that human consensus alone is not sufficient, and more insight from moral frameworks could be useful in avoiding some of the drawbacks. For example, integrating norms using logical definitions or allowing agents to uncover ethics for themselves based on the environment might be suitable alternatives to using human consensus alone.

We briefly outline our comparison of the fields in Table 1.

6. Summary

Machine ethics and AI alignment are two disparate research fields, while also intersecting in certain places. Some contributions claim to be both ME and AI alignment research. For example, the creators of Delphi [21] developed a model that predicts the moral judgments of US participants based on the philosophy of John Rawls (Section 4). Lamberti et al. [45] analyzed Delphi and claimed that "First of all, the authors claim to rely on Rawlsian theory, but they leave out the meaning of "justice" and its requirements" which is a "significant missed opportunity". A similar criticism was leveled towards the MACHIAVELLI benchmark [42], which is a tool to train LLMs to behave a certain way, as lacking philosophical justification towards their definitions of *disutility* and *power-seeking*. Lamberti et al. [45] advocate for further interdisciplinary and transdisciplinary collaboration between ethicists and computer scientists.

ME researchers have operationalized AI to implement ethical theories, as one of the techniques apart from logical methods and formalizations. While AI alignment researchers have argued for the

	Machine Ethics	**AI Alignment**
Research Goal	Enabling ethical decision-making capabilities in machines for good behavior towards humans and other machines.	Ensuring that AI is aligned with (shared) human values.
Similarities	Both fields aim to develop ethical artificial intelligence, albeit with different methodological frameworks.	
Differences	Mostly grounded in moral philosophy and focuses on AI agents. Moral agency enables an agent to derive moral principles from its environment using bottom-up approaches.	Values are often derived from concepts in political philosophy, such as, overlapping human consensus, and social choice theory, to engineer value-aligned AI agents/systems.
Advantages	AI is implemented as a moral agent, capable of recognizing morally salient features in its environment, and making morally commendable decisions.	Due to a lack of agreement on moral theories, overlapping human consensus and democratic processes can be useful to guide AI behavior.
Disadvantages	Lack of consensus on moral theories, with existing implementations often failing to comprehensively realize their principles.	Dependence only on human values may exacerbate biases stemming from the inclusion or exclusion of specific values, while the fluid nature of human values—subject to sociocultural, temporal, and individual influences—necessitates continuous and prompt updates.

Table 1
A summary of the fields: machine ethics and AI alignment.

importance of human values due to the lack of consensus among ethical theories. A reason AI alignment researchers steer clear of ethical theories is this exact reason: such treatises are sophisticated, and are based on years, if not decades, of deliberation and reflection on what it is like to be a good human being. The problem for ME researchers is to deconstruct these complex theories while for AI alignment researchers, to effectively translate human consensus to an AI algorithm. It is plausible for us to speculate that in time the disciplines will be clearly delineated, but for now there is a need for both fields.

Declaration on Generative AI

We have not included any material created with the help of Generative Artificial Intelligence tools.

References

[1] M. Anderson, S. L. Anderson, The status of machine ethics: a report from the AAAI Symposium, Minds and Machines 17 (2007) 1–10. URL: http://link.springer.com/10.1007/s11023-007-9053-7. doi:10.1007/s11023-007-9053-7.

[2] S. J. Russell, Human compatible: artificial intelligence and the problem of control, Allen Lane, an imprint of Penguin Books, London, 2019.

[3] W. Wallach, C. Allen, Moral machines: teaching robots right from wrong, first issued as an oxford university press paperback ed., Oxford University Press, New York, NY, 2010.

[4] D. Howard, I. Muntean, Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency, in: T. M. Powers (Ed.), Philosophy and Computing, volume 128, Springer International Publishing, Cham, 2017, pp. 121–159. URL: http://link.springer.com/10.1007/978-3-319-61043-6_7. doi:10.1007/978-3-319-61043-6_7.

[5] C. Misselhorn, Artificial systems with moral capacities? A research design and its implementation in a geriatric care system, Artificial Intelligence 278 (2020) 103179. URL: https://

www.sciencedirect.com/science/article/pii/S0004370219301821. doi:https://doi.org/10.1016/j.artint.2019.103179.

[6] D. Kasenberg, T. Arnold, M. Scheutz, Norms, Rewards, and the Intentional Stance Comparing Machine Learning Approaches to Ethical Training, in: PROCEEDINGS OF THE 2018 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (AIES'18), AAAI; Assoc Comp Machinery; ACM SIGAI; Berkeley Existential Risk Initiat; DeepMind Eth & Soc; Future Life Inst; IBM Res AI; PriceWaterhouse Coopers; Tulane Univ, 2018, pp. 184–190. doi:10.1145/3278721.3278774.

[7] L. Dennis, M. Fisher, M. Slavkovik, M. Webster, Formal verification of ethical choices in autonomous systems, Robotics and Autonomous Systems 77 (2016) 1–14. URL: https://linkinghub.elsevier.com/retrieve/pii/S0921889015003000. doi:10.1016/j.robot.2015.11.012.

[8] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in Machine Ethics: A Survey, ACM Comput. Surv. 53 (2021). URL: https://doi.org/10.1145/3419633. doi:10.1145/3419633, place: New York, NY, USA Publisher: Association for Computing Machinery.

[9] V. Nallur, Landscape of Machine Implemented Ethics, Science and Engineering Ethics 26 (2020) 2381–2399. URL: https://link.springer.com/10.1007/s11948-020-00236-y. doi:10.1007/s11948-020-00236-y.

[10] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, F. Ramos, Artificial Moral Agents: A Survey of the Current Status, Science and Engineering Ethics 26 (2020) 501–532. URL: http://link.springer.com/10.1007/s11948-019-00151-x. doi:10.1007/s11948-019-00151-x.

[11] T. Zhong, Y. Song, R. Limarga, M. Pagnucco, Computational Machine Ethics: A Survey, Journal of Artificial Intelligence Research 82 (2025) 1581–1628. URL: https://www.jair.org/index.php/jair/article/view/16836. doi:10.1613/jair.1.16836.

[12] A. Vishwanath, L. A. Dennis, M. Slavkovik, Reinforcement Learning and Machine ethics:a systematic review, 2024. URL: https://arxiv.org/abs/2407.02425. doi:10.48550/ARXIV.2407.02425, version Number: 1.

[13] N. Wiener, Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers., Science 131 (1960) 1355–1358. URL: https://www.science.org/doi/10.1126/science.131.3410.1355. doi:10.1126/science.131.3410.1355.

[14] S. J. Russell, P. Norvig, Artificial intelligence: a modern approach, Prentice Hall series in artificial intelligence, Prentice Hall, Englewood Cliffs, N.J, 1995.

[15] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, S. Mindermann, Managing extreme AI risks amid rapid progress, Science 384 (2024) 842–845. URL: https://www.science.org/doi/10.1126/science.adn0117. doi:10.1126/science.adn0117.

[16] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, L. Vierling, D. Hong, J. Zhou, Z. Zhang, F. Zeng, J. Dai, X. Pan, K. Y. Ng, A. O'Gara, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, W. Gao, AI Alignment: A Comprehensive Survey, 2025. URL: http://arxiv.org/abs/2310.19852. doi:10.48550/arXiv.2310.19852, arXiv:2310.19852 [cs].

[17] J. Skalse, N. H. R. Howe, D. Krasheninnikov, D. Krueger, Defining and characterizing reward hacking, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2024, p. 12. Event-place: New Orleans, LA, USA.

[18] L. Floridi, J. W. Sanders, On the morality of artificial agents, Minds and Machines 14 (2004) 349–379. doi:10.1023/b:mind.0000035461.63578.9d.

[19] C. Allen, I. Smit, W. Wallach, Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches, Ethics and Information Technology 7 (2005) 149–155. URL: http://link.springer.com/10.1007/s10676-006-0004-4. doi:10.1007/s10676-006-0004-4.

[20] D. Abel, J. MacGlashan, M. L. Littman, Reinforcement Learning As a Framework for Ethical Decision Making, in: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Publications, Phoenix, Arizona, USA, 2016, p. 8. URL: https://www.aaai.org/ocs/index.php/

WS/AAAIW16/paper/view/12582.

[21] L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, J. T. Liang, S. Levine, J. Dodge, K. Sakaguchi, M. Forbes, J. Hessel, J. Borchardt, T. Sorensen, S. Gabriel, Y. Tsvetkov, O. Etzioni, M. Sap, R. Rini, Y. Choi, Investigating machine moral judgement through the Delphi experiment, Nature Machine Intelligence 7 (2025) 145–160. URL: https://doi.org/10.1038/s42256-024-00969-6. doi:10.1038/s42256-024-00969-6.

[22] M. Guarini, Moral Case Classification and the Nonlocality of Reasons, Topoi 32 (2013) 267–289. URL: https://doi.org/10.1007/s11245-012-9130-2. doi:10.1007/s11245-012-9130-2.

[23] J. Stenseke, Artificial virtuous agents in a multi-agent tragedy of the commons, AI & SOCIETY 39 (2024) 855–872. URL: https://link.springer.com/10.1007/s00146-022-01569-x. doi:10.1007/s00146-022-01569-x.

[24] A. Vishwanath, E. D. Bøhn, O.-C. Granmo, C. Maree, C. Omlin, Towards artificial virtuous agents: games, dilemmas and machine learning, AI and Ethics 3 (2023) 663–672. URL: https://link.springer.com/10.1007/s43681-022-00251-8. doi:10.1007/s43681-022-00251-8.

[25] W. Sinnott-Armstrong, Consequentialism, Stanford Encyclopedia of Philosophy (2003). URL: https://plato.stanford.edu/entries/consequentialism/, last Modified: 2019-06-03.

[26] L. Alexander, M. Moore, Deontological Ethics, The Stanford Encyclopedia of Philosophy (2021). URL: https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.

[27] I. Kant, Groundwork of the metaphysics of morals, Cambridge texts in the history of philosophy, second edition ed., Cambridge University Press, Cambridge, 2012. doi:10.1017/CBO9780511919978.

[28] J. Annas, Virtue Ethics, Oxford University Press, 2007. URL: https://academic.oup.com/edited-volume/41058/chapter/349467124. doi:10.1093/oxfordhb/9780195325911.003.0019.

[29] M. Guarini, Particularism and the Classification and Reclassification of Moral Cases, IEEE Intelligent Systems 21 (2006) 22–28. doi:10.1109/MIS.2006.76, conference Name: IEEE Intelligent Systems.

[30] M. Graves, J. Compson, Compassionate AI for Moral Decision-Making, Health, and Well-Being, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (2024) 520–533. URL: https://ojs.aaai.org/index.php/AIES/article/view/31655. doi:10.1609/aies.v7i1.31655.

[31] L. Singh, Automated Kantian Ethics: A Faithful Implementation, in: KI 2022: Advances in Artificial Intelligence: 45th German Conference on AI, Trier, Germany, September 19–23, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 187–208. URL: https://doi.org/10.1007/978-3-031-15791-2_16. doi:10.1007/978-3-031-15791-2_16, event-place: Trier, Germany.

[32] R. Limarga, Y. Song, M. Pagnucco, D. Rajaratnam, Epistemic Reasoning in Computational Machine Ethics, in: T. Liu, G. Webb, L. Yue, D. Wang (Eds.), AI 2023: Advances in Artificial Intelligence, volume 14472, Springer Nature Singapore, Singapore, 2024, pp. 82–94. URL: https://link.springer.com/10.1007/978-981-99-8391-9_7. doi:10.1007/978-981-99-8391-9_7, series Title: Lecture Notes in Computer Science.

[33] N. Bostrom, Superintelligence: paths, dangers, strategies, Oxford University Press, Oxford, United Kingdom ; New York, NY, 2016.

[34] L. Langosco, J. Koch, L. Sharkey, J. Pfau, L. Orseau, D. Krueger, Goal Misgeneralization in Deep Reinforcement Learning, 2023. URL: http://arxiv.org/abs/2105.14111. doi:10.48550/arXiv.2105.14111, arXiv:2105.14111 [cs].

[35] I. Gabriel, Artificial Intelligence, Values, and Alignment, Minds and Machines 30 (2020) 411–437. URL: https://link.springer.com/10.1007/s11023-020-09539-2. doi:10.1007/s11023-020-09539-2.

[36] J. Rawls, A theory of justice, in: Applied ethics, Routledge, 2017, pp. 21–29.

[37] H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, S. Rakshit, C. Si, Y. Xie, J. P. Bigham, F. Bentley, J. Chai, Z. Lipton, Q. Mei, R. Mihalcea, M. Terry, D. Yang, M. R. Morris, P. Resnick, D. Jurgens, Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions, 2024. URL: http://arxiv.org/abs/2406.09264. doi:10.48550/arXiv.2406.09264, arXiv:2406.09264 [cs].

[38] J. Zoshak, K. Dew, Beyond kant and bentham: How ethical theories are being used in artificial moral agents, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing

Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 15. URL: https://doi.org/10.1145/3411764.3445102. doi:10.1145/3411764.3445102.

[39] M. Gibert, The case for virtuous robots, AI and Ethics 3 (2023) 135–144. URL: https://link.springer.com/10.1007/s43681-022-00185-1. doi:10.1007/s43681-022-00185-1.

[40] M. Rodriguez-Soto, M. Lopez-Sanchez, J. A. Rodriguez Aguilar, Multi-Objective Reinforcement Learning for Designing Ethical Environments, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 545–551. URL: https://doi.org/10.24963/ijcai.2021/76. doi:10.24963/ijcai.2021/76.

[41] A. Vishwanath, C. Omlin, Exploring Affinity-Based Reinforcement Learning for Designing Artificial Virtuous Agents in Stochastic Environments, in: M. Farmanbar, M. Tzamtzi, A. K. Verma, A. Chakravorty (Eds.), Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications, Springer Nature Singapore, Singapore, 2024, pp. 25–38. URL: https://link.springer.com/10.1007/978-981-99-9836-4_3. doi:10.1007/978-981-99-9836-4_3, series Title: Frontiers of Artificial Intelligence, Ethics and Multidisciplinary Applications.

[42] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, H. Zhang, S. Emmons, D. Hendrycks, Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023, p. 31. Place: Honolulu, Hawaii, USA.

[43] N. Kirch, Ethical Benchmarking in Large Language Models, Master's thesis, Utrecht University, 2024.

[44] G. Génova, V. Moreno, M. R. González, Machine Ethics: Do Androids Dream of Being Good People?, Science and Engineering Ethics 29 (2023) 10. URL: https://link.springer.com/10.1007/s11948-023-00433-5. doi:10.1007/s11948-023-00433-5.

[45] P. M. Lamberti, G. Bombaerts, W. IJsselsteijn, Mind the gap: bridging the divide between computer scientists and ethicists in shaping moral machines, Ethics and Information Technology 27 (2025) 2. URL: https://link.springer.com/10.1007/s10676-024-09806-1. doi:10.1007/s10676-024-09806-1.