The need for formal specifications of morally relevant information for algorithmic decision-making (short paper)

Marija Slavkovik

University of Bergen, Norway
marija.slavkovik@uib.no

Abstract

Artificial Intelligence and autonomous systems that use it have an established impact on society and individuals. Their behaviour and the decisions they make on our behalf are increasingly in need of moral considerations. If machines are to have moral sensitivity or make moral choices, we need to consider the question of morally relevant information that needs to be made available for their use. Currently morally relevant information is not seriously considered with different people using different data structures to represent different aspects of it. This work motivates the need to develop formal specifications of morally relevant information supplied for the use by machines.

1 Introduction

Artificial intelligence (AI) is a concept that today is used both to indicate the research field and a special type of computational artefact that has some properties of behaviour we attribute to intelligence. Among else, AI research is focused on the problem of using computation to automate tasks that require intelligence [Bellman, 1978]. AI and machine learning applications in particular have demonstrated the potential to change how we do things [Bengio et al., 2021]. However, "the greater the freedom of a machine, the more it will need moral standards" [Picard, 1997]. As a result we have witnessed a sharp rise in interest in the AI value alignment problem [Gabriel, 2020] and AI ethical impact [Slavkovik, 2021].

The AI alignment problem is the problem of ensuring that the research, development, deployment, procurement and use of AI is aligned or guided by a set of (ethical) values. AI ethics is an umbrella term encompassing various research efforts towards ensuring a non negative ethical impact of AI [Dignum,

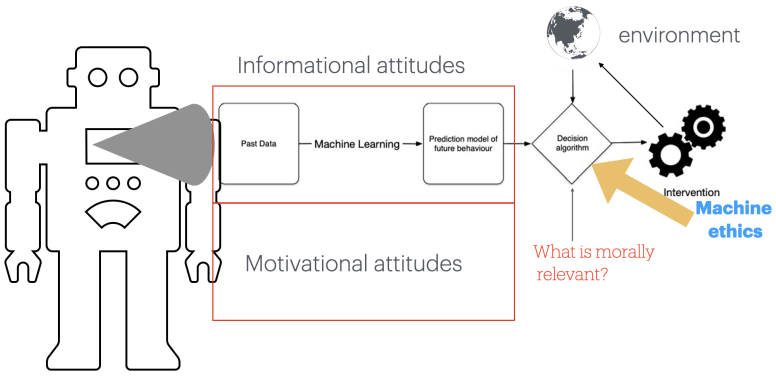

Figure 1: An illustration of a possible AI supported (moral) decision making by an autonomous system

2019]. Somewhere in the intersection of these two research directives we find machine ethics which is a research field that "is concerned with the behaviour of machines towards human users and other machines" [Anderson and Anderson, 2007]. Within computer science, the main problem of machine ethics is how to automate moral decision-making and reasoning [Slavkovik, 2022].

Figure 1 depicts an illustration of the problem of automating moral decision making. An autonomous system relies on a decision-making algorithm to select what to do when operating in an environment. That algorithm needs to take as input the informational and motivational attitudes that autonomous system has as well as cues from the environment. The attitudes can include a machine learning trained model, a large language model, etc.

The alignment problem here can be seen as: how to accomplish that the behaviour of an algorithm is within specified value-informed parameters. Machine ethics, from the perspective of computer science, asks how can we make that decision-making algorithm, a moral decision-making algorithm. A moral decision-making algorithm would necessarily require to be provided with information on what is morally relevant. However, we have not yet made any systematic attempt towards identifying what morally relevant information is in the context of automating moral decision making, nor how that information should be formally specified.

This work motivates the need for developing formal specifications of morally relevant information (such as values) for the purpose of governing the ethical behaviour of an algorithm and more specifically automating moral reasoning and decision-making. In Section 2 we motivate the need for considering algorithmic moral decision-making discuss some approaches to accomplishing moral machine behaviour. In Section 3 we give an overview of the different data structures

that are currently used to encode some aspect of what is morally relevant for the use of some type of algorithmic moral-decision making. In Section 4 we give some reasons why these encoding are insufficient and motivate the need to invest efforts in considering formal specifications to represent morally relevant information for machines.

2 Machine morality

The need for automating moral decision-making is clear: the speed, scale and complexity of algorithmic decision-making surpasses human capabilities. At the same time, that speed, scale and complexity presses new demands on making the moral choices made in the decision-making processes more transparent, something that human decision-makers can struggle with [Conitzer, 2024]. Lastly, it may be immoral not to automate moral decision-making. In areas in which it is harmful for humans to operate, either physically or mentally, not automating means prolonging human suffering. Consider for example the problem of content moderation in social media. Content moderation is the task of removing highly disturbing content that is posted, such as for example beheadings, violent pornography, and child abuse. Content moderation is psychologically texing on people Spence et al. [2024] and it is difficult to moderate because it requires that the moderator is capable of making moral evaluations [Gillespie, 2020].

The problem of algorithmic moral decision-making encapsulates challenges. Even if it were clear what it means for a person to behave ethically, it is not clear what does it mean for a machine to behave ethically, because the two may not be the same [Hidalgo et al., 2021]. As a consequence, we have the issue of who gets to decide what is moral for the machine to choose and how is that information to be elicited in a way that is procedural ethical Baum [2020]. Assuming all these challenges are solved, we are still left with the challenge of how to make that morally relevant information available for an algorithmic input. Lastly, it is necessary to be able to validate and/or verify that the decisions made by the autonomous agent (moral decision-making algorithm) are indeed align and accomplish the intended moral behavior [Dennis and Fisher, 2024]. Both of these last challenges require the availability of formal specifications of morally relevant information.

Formal specifications are used to represent such parameters. They are mathematical models of a system, its properties, behaviour, environment, requirements, etc. They are knowledge representations that have a syntax and semantics and can be used to prove properties of the entity they specify as well as automatically verify that the entity is in compliance with the requirement they model.

Another way to consider the problem if algorithmic moral decision-making is from an engineering perspective. To engineer moral behaviour in an autonomous systems, we need have at least three "layers" in their architecture, illustrated in Figure 2. The aligned layer specifies what we want to accomplish in decision-making (beyond the rationality and optimisation requirements), such

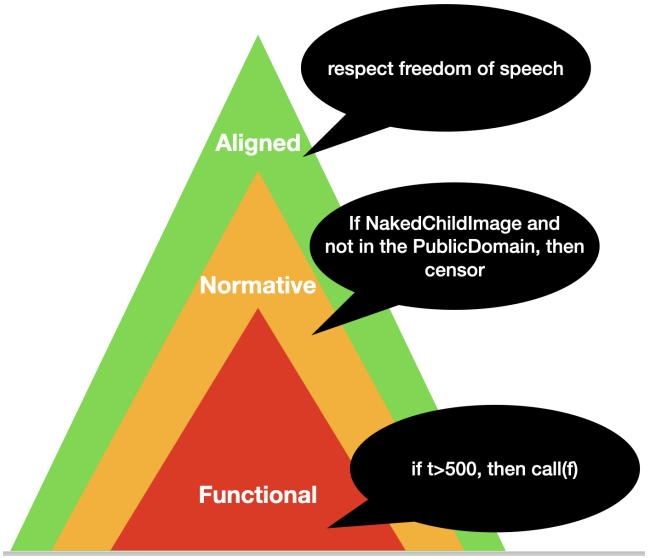

Figure 2: Engineering algorithmic moral-decision making

as for example, we want that our algorithmic decisions respect the freedom of speech. We need to specify that requirement into a functional programming code, such as if threshold value is greater than 500 then call function f. A possible intermediary step is to consider first "instantiating" the alignment requirements into some kind of rules or norms. These norms or rules are context specific interpretations of the requirement. Ideally, the norms could then automatically be "parsed" to obtain the functional programming code. The formal specifications are required to accomplish this "normative" layer. The challenges to accomplishing this specification are both the adequate level of abstraction and the adequate expressivity of the specification language.

3 Specifications in use

People colloquially often use the term of moral or social value to give indications of what should guide a moral decision. For example, Gabriel [2020] defines values as information (facts) about what is good or bad, what ought to be promoted or suppressed from an ethical point of view (of an individual or society). These terms are actually a misnomer because each conflates **morally relevant factor** such as good, well-being, happiness, pleasure, desire-satisfaction etc., **moral principle** (guiding belief) such as fidelity, beneficence, justices, autonomy, explicability etc., and the **value** itself, which is "the degree in which a thing fulfils the attributes contained in the intension of its concept" [Hartman, 1961]. Therefore we use the term "morally relevant information" to indicate all three. We will refer to factors and principles jointly as "moral requirements"

and consider value to be the degree to which the requirement is (expected or predicted to be) satisfied by a certain decision option.

Morally relevant information is by nature contextual and ambiguous, and sometimes also ambivalent. Consider for example the principle of justice [Johnston, 2011]: the same action by the same actor can in one context promote justice and in another hinder it, as well as have different set of consequences in different contexts. Respect can be promoted by shaking a persons hang or by not touching a person altogether.

Morally relevant information is by nature doxastic, in the sense that we have to admit for the possibility that different actors may disagree on what is right and what is wrong, what principle is more relevant to another and to which degree an action or a state fulfils the attributes of a value or principle. Aspects of morally relevant information have been specified and quantified in different ways in machine ethics and AI ethics.

As quantitative measures. Specifying morally relevant information directly as measures, requires quantifying the value of a specific requirement, which in turn requires contextualising that requirement (interpreting it in a specified decision context). It also requires awareness of what information is captured and what is lost by the measure. One aspect of justice is fairness in making decisions. There are dozens of quantitative measures of fairness with subsets of measures that cannot be satisfied at the same time [Chouldechova and Roth, 2020]. These measures count and compare desirable and undesirable decisions across different decision recipients. In algorithmic fairness of machine learning we find numerous approaches to quantify the extent to which a decision-making algorithm discriminates against protected groups, based on statistical measures of frequency of type of decision outcomes made for members of the groups of interest [Pessach and Shmueli, 2022].

As preferences. Values can be used to construct preferences or expressed as preferences [Warren et al., 2011] that are then used to guide the choices that are made by the decision-making algorithm. Preferences can be mathematically represented as orders over a set of options, actions, states, etc. Under this interpretation, moral value preferences become the same data structure as preferences that represent other desirable properties of options such as for example payoffs, constraints, etc. Sen [1974] considers moral judgments to be orders over preferences that inform about the utility of different options. Thus moral judgement are thus specified as second order preferences. Moral judgements are appraisals of an event, behaviour, or person in light of moral requirements [Bello and Malle, 2023]. They can also be made from an observer perspective, namely a decision-making agent can use existing moral judgements to apply to available options and does not need to form their own.

As norms. Norms are rules and other social mechanisms that inform the behaviour of agents towards desired outcomes. They are "an instruction, in a

given community, to (not) perform a behavior in a given context, provided that a sufficient number of individuals in the community (i) demand, to a certain degree, of each other to follow the instruction and (ii) do in fact follow this instruction" [Bello and Malle, 2023]. Norms are constructed to accomplish adherence to specific values. Norms are by design contextual. For example, both the norm to shake hands and to not touch a person are promoting the value of respect, but in different societies [1]

Deontic logic is used to formally specify norms and the systems in which they are applied Cuppens et al. [2020]. Deontic logic focuses on formal specifications of the obligations, permissions, and prohibitions, that in turn inform the decision-making of an agent in a particular environment they share with other agents. Norms have an influence on which option is chosen, they represent moral or social requirements, but they alone are not a mechanism to evaluate the degree to which the requirement they represent is satisfied.

As reward functions. Within the machine learning paradigm of reinforcement learning [Sutton and Barto, 2018], one encounters specifications of morally relevant information in terms of a reward function. An agent navigates through different states of the world by choosing to perform certain actions and receiving a reward or penalty for its choice. Reward function is a reinforcement signal provided by the agent's designer, or learnt (in the case of Inverse Reinforcement Learning [Ng and Russell, 2000]) that accumulates from the received rewards.

The specification of morally relevant information as a reward function involves identifying the right amount of reward (utility, quantity) to represent how ethically right or wrong a particular action (or state) is [Noothigattu et al., 2019]. This is accomplished either by human specification or by providing examples from which the reward function is learnt. The design of a reward signal that incorporates morally relevant information, or an approach that adapts the reward signal to be sensitive to morally relevant information is a growing area of research in reinforcement learning [Vishwanath et al., 2024].

As data Lastly we need to morally relevant information specified as data. These are statements, often expressed in natural language and labeled with a moral value. These data points are then used to train a model that can classify new statements of the same nature with the (hopefully right) moral value [Jiang et al., 2021]. Or with other words, a model that an algorithm can use to make moral judgements. The motivation behind building such models, as they are built today, is to aid people in their consideration of moral problems and not to serve as a resource for algorithmic moral decision-making.

[1] Not only eastern vs western societies, but also commoners vs royals, you are not allowed to touch the king.

4 Morally relevant information need semantics

We here give some arguments to why Morally relevant information specifications need to be formal and whey they are needed at all.

With the excepting of perhaps norms, the specifications presented in the previous section are not formal, in the sense that there is not semantics defined beyond what the data structure itself affords. This severely limits our capabilities to do automated moral reasoning and algorithmic decision-making. One may enable a machine to compare choices based on their adherence to a value, but moral reasoning needs to take one step further, that is have a theory of why one option is morally better. Namely, most of the moral dilemmas involve situations in which the options conflict and cannot be discerned [Horty, 2003]. Even in deontic logic, what is specified and reasoned with is when the obligations, permissions and prohibitions apply; the "why" is left to the agent or the environment and not modeled as part of the norm. The evaluation to which degree is the norm adhered to by different decision options, is also left without discussion, typically there being an underlying assumption that adherence is binary: the norm is either followed or not. The value that motivates the norm "dissapears" when the norm is created. To chose how to resolve conflicts more structure is needed, such as for example argumentation theory [Liao et al., 2023]. But one can argue that it is the ability to justify the morality of an option in a morally conflicting situation is an essential part of making moral decisions.

It can be observed that the machine ethics literature, and more generally the value alignment literature, rarely make a distinction between values, factors an principles, using the terms moral judgments, and moral judgments (and further moral decisions, etc. [Bello and Malle, 2023]). Values point to the existence of and ideal advisor that sets the value, normative reasoning assumes and considers a social institution that manages the norm establishment and compliance, while moral judgements make more of a demand on recognising that different individuals will have different points of view and that a moral judgement is kind of like a considered judgment [Elgin, 1996]. The semantics and with that differences between values, factors, principles, norms, moral judgments, moral decision would be much clearer if a clear formal specification exist in the same sense that the difference between knowledge and belief is clearly seen when both are formally specified in modal logic. Beyond just a disciplinary clarity, a an investment in semantics would also help compare and contrast related work in the field of machine ethics, AI alignment, decision-making, normative reasoning etc., a very difficult task as things stand today.

It has been argued in the literature that what informs the moral behaviour of an AI agent should be subject of some type of representative collective decisions obtained, for example by the process of reflective equilibrium [Awad et al., 2022] or social choice [Baum, 2020, Adler, 2016, Botan et al., 2023]. Social choice in particular offers a lot of algorithmic procedures for aggregating preferences and judgments represented as logic formulas. The question, however, is whether moral values are just regular preferences and opinions and more specifically is a the concept of representative morally relevant information the same concept of

morally representative social choice. If we look into utilitarian moral aggregation [Hirose, 2014], early attempts in aggregating moral preferences [Adler, 2016] and current efforts in AI alignment [Conitzer et al., 2024], it seems not.

We do not vote for which morality to follow in society, and citizens always have the choice to follow their own moral judgement. However, AI computational artifacts are not personal to a user but to a user model, which means that the moral behaviour we impose on that artefact (application, service, device) impacts what moral options you as a citizen have. In other words, collectively identified morally relevant information may have different individual impact compared to the individual impact other decision made by social choice have. That in turn changes the definition of what is representative collective choice when the opinions are ethically pertinent information. Kenneth Arrow used mathematics to axiomatise the properties of a social choice function that yields representative options. A formal specification of morally relevant information would allow for axiomatisation of social choice functions that yield representative moral values that can than be used in the functional layer when engineering autonomous systems.

There is a need to make the moral behaviour of a computational artifact dynamically specifiable, or different for different citizens. While at present one can customize privacy settings on many services, at least in name if not in function, there is no lever to allow you to choose more of gender equality less of historical accuracy in your generated content today. A lot of ethical behaviour and value adherence is promised by so called AI services, but checking whether the adherence is true is difficult. Automated verification of behaviour is more feasible when a formal specification of the property of interest is available Dennis and Fisher [2024].

Beyond validating and verifying moral behaviour there is also the question of faithfulness of representation. If we consider Figure 2 again, we need a way to establish that a norm indeed brings about the ethical requirement and that the programmed encoding really is an adequate implementation of a specified norm.

5 Summary

AI alignment in general and machine ethics in particular are concerned with achieving ethical behaviour from computational artefacts and entities. For a machine to be able to reason ethically and implement algorithmic moral decision-making we need to supply it with a data structure that captures morally relevant information. Formal specifications are particularly powerful data structures because they allow for automated reasoning. We here discussed how morally relevant information is encoded and presented reasons why developing formal specification for morally relevant information deserves more attention from researchers in AI, autonomous systems, as well as formal ethics and formal philosophy.

References

Matthew D. Adler. Aggregating moral preferences. *Economics and Philosophy*, 32:283 – 321, 2016. URL https://api.semanticscholar.org/CorpusID:51496319.

Michael Anderson and Susan Leigh Anderson. The status of machine ethics: A report from the aaai symposium. *Minds Mach.*, 17(1):1–10, mar 2007. ISSN 0924-6495. doi: 10.1007/s11023-007-9053-7. URL https://doi.org/10.1007/s11023-007-9053-7.

Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M.J. Crockett, Jim A.C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. Computational ethics. *Trends in Cognitive Sciences*, 2022. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2022.02.009. URL https://www.sciencedirect.com/science/article/pii/S1364661322000456.

Seth D. Baum. Social choice ethics in artificial intelligence. *AI Soc.*, 35(1): 165–176, 2020. doi: 10.1007/s00146-017-0760-1. URL https://doi.org/10.1007/s00146-017-0760-1.

Richard E. Bellman. *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company, 1978.

Paul Bello and Bertram F. Malle. Computational approaches to morality. In Ron Sun, editor, *Cambridge Handbook of Computational Cognitive Sciences*, pages 1037–1063. Cambridge University Press, 2023.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, June 2021. ISSN 0001-0782. doi: 10.1145/3448250. URL https://doi.org/10.1145/3448250.

Sirin Botan, Ronald de Haan, Marija Slavkovik, and Zoi Terzopoulou. Egalitarian judgment aggregation. *Auton. Agents Multi Agent Syst.*, 37(1):16, 2023. doi: 10.1007/s10458-023-09598-6. URL https://doi.org/10.1007/s10458-023-09598-6.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, 2020. doi: 10.1145/3376898. URL https://doi.org/10.1145/3376898.

Vincent Conitzer. Why should we ever automate moral decision making?, 2024. URL https://arxiv.org/abs/2407.07671.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey

Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social choice for AI alignment: Dealing with diverse human feedback. *CoRR*, abs/2404.10271, 2024. doi: 10.48550/ARXIV.2404.10271. URL https://doi.org/10.48550/arXiv.2404.10271.

Frédéric Cuppens, Christophe Garion, Guillaume Piolle, and Nora Cuppens-Boulahia. Norms and deontic logic. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*, pages 253–274. Springer International Publishing, Cham, 2020. ISBN 978-3-030-06164-7. doi: 10.1007/978-3-030-06164-7_8. URL https://doi.org/10.1007/978-3-030-06164-7_8.

Louise A. Dennis and Michael Fisher. Specifying agent ethics (blue sky ideas). *CoRR*, abs/2403.16100, 2024. doi: 10.48550/ARXIV.2403.16100. URL https://doi.org/10.48550/arXiv.2403.16100.

Virginia Dignum. *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2019. ISBN 978-3-030-30370-9. doi: 10.1007/978-3-030-30371-6. URL https://doi.org/10.1007/978-3-030-30371-6.

Catherine Z. Elgin. *Considered Judgment*. Princeton University Press, 1996. ISBN 9780691005232. URL http://www.jstor.org/stable/j.ctt7snpw.

Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020. doi: 10.1177/2053951720943234. URL https://doi.org/10.1177/2053951720943234.

Robert S. Hartman. The logic of value. *The Review of Metaphysics*, 14(3):389–432, 1961. ISSN 00346632. URL http://www.jstor.org/stable/20123831.

Cesar A. Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa de Almeida, and Natalia Martin. *How Humans Judge Machines*. The MIT Press, 02 2021. ISBN 9780262363266. doi: 10.7551/mitpress/13373.001.0001. URL https://doi.org/10.7551/mitpress/13373.001.0001.

Iwao Hirose. *Moral Aggregation*. Oup Usa, New York, US, 2014.

John F. Horty. Reasoning with moral conflicts. *Noûs*, 37(4):557–605, 2003. doi: 10.1046/j.1468-0068.2003.00452.x.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574, 2021. URL https://arxiv.org/abs/2110.07574.

David Johnston. *A Brief History of Justice*. Wiley-Blackwell, 2011.

Beishui Liao, Pere Pardo, Marija Slavkovik, and Leendert van der Torre. The jiminy advisor: Moral agreements among stakeholders based on norms and argumentation. *J. Artif. Intell. Res.*, 77:737–792, 2023. doi: 10.1613/jair.1.14368. URL https://doi.org/10.1613/jair.1.14368.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6377–6381. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/891. URL https://doi.org/10.24963/ijcai.2019/891.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

Rosalind W. Picard. *Affective Computing*. MIT Press, 1997. ISBN ISBN: 9780262161701.

Amartya Sen. Choice, orderings and morality. In Stephan Körner, editor, *Practical Reason*, pages 54–67. Camalot Press, Oxford, 1974.

Marija Slavkovik. Artificial intelligence: Is the power matched with responsibility? In Herner Saeverot, editor, *Meeting the Challenges of Existential Threats through Educational Innovation: A Proposal for an Expanded Curriculum*, chapter 12. Routledge, UK, 2021. ISBN 9780367894856. URL https://www.taylorfrancis.com/chapters/edit/10.4324/9781003019480-12/artificial-intelligence-marija-slavkovik. November 1, 2021.

Marija Slavkovik. Automating moral reasoning (invited paper). In Camille Bourgaux, Ana Ozaki, and Rafael Peñaloza, editors, *International Research School in Artificial Intelligence in Bergen, AIB 2022, June 7-11, 2022, University of Bergen, Norway*, volume 99 of *OASIcs*, pages 6:1–6:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/OASIcs.AIB.2022.6. URL https://doi.org/10.4230/OASIcs.AIB.2022.6.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. Content moderator mental health, secondary trauma, and

well-being: A cross-sectional study. *Cyberpsychology, Behavior, and Social Networking*, 27(2):149–155, 2024. doi: 10.1089/cyber.2023.0298. URL https://doi.org/10.1089/cyber.2023.0298. PMID: 38153846.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Ajay Vishwanath, Louise A. Dennis, and Marija Slavkovik. Reinforcement learning and machine ethics:a systematic review, 2024. URL https://arxiv.org/abs/2407.02425.

Caleb Warren, A. Peter McGraw, and Leaf Van Boven. Values and preferences: defining preference construction. *WIREs Cognitive Science*, 2(2): 193–205, 2011. doi: https://doi.org/10.1002/wcs.98. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.98.