

# Moral Exercises for Human Oversight of Algorithmic Decision-Making

Teresa Scantamburlo<sup>1,2,\*</sup>, Silvia Crafa<sup>3</sup> and Giovanni Grandi<sup>4</sup>

<sup>1</sup>Ca' Foscari University of Venice, via Torino 155, 30172 Venice, Italy

<sup>2</sup>European Centre for Living Technology, Dorsoduro 3911, Calle Crosera, 30123 Venice, Italy

<sup>3</sup>University of Padua, Via Trieste 63, 35121 Padua, Italy

<sup>4</sup>University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy

## Abstract

In this paper we present an ethical framework aimed at supporting human agency and oversight throughout the life-cycle of algorithmic decision-making systems. Drawing upon classical philosophical traditions, we shift the focus from ethical solutions to the role of Artificial Intelligence (AI) actors in a quality decision process. To this aim the framework highlights the dynamic nature of people's moral decisions, and the "ethical tools" that are inherent within the human being. The primary objective is not to enforce morality within the *machine* itself but to cultivate moral agency in the *human*. This offers the conceptual coordinates to put forward a set of "moral exercises", practical activities that can be used for the moral training of human actors involved in the life process of AI-based decision systems. Rather than being algorithmic procedures or workflows for ensuring "moral outcomes", these exercises are flexible instruments to shape the human processes underlying the oversight of AI systems. We illustrate the practical implications of our framework by showing potential cases of application of the exercises, and by creating connections with existing AI ethics methodologies.

## Keywords

Human Oversight, Moral Exercise, Responsible AI, AI ethics, Algorithmic decision-making

## 1. Introduction

The widespread concern upon the proliferation of abstract ethical principles in the field of Artificial Intelligence (AI) has spurred the development of a number of tools for ethical assessment and auditing, aimed at offering concrete solutions [1]. However, the efforts to close the gap between principles and practices distracted us from more radical questions about the meaning of ethics in the context of AI innovation.

In particular, insisting on ethical solutions may leave for granted that dealing with the ethics of AI means, first and foremost, to implement a procedure or follow a specific workflow. In this paper we want to engage with more fundamental questions such as: what does it mean acting ethically in the context of AI? What does responsible behavior imply for AI innovation? In


---

22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), November 6 - 9, 2023, Rome, Italy

\*Corresponding author.

✉ teresa.scantamburlo@unive.it (T. Scantamburlo); crafa@math.unipd.it (S. Crafa); giovanni.grandi@units.it (G. Grandi)

ORCID 0000-0002-3769-8874 (T. Scantamburlo)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

other terms, we shift the focus from the production of ethical outcomes to the role of people in a quality decision-making process.

To this aim we propose a framework that highlights key ethical dynamics in decision-making processes (at both the individual and group levels) and inspires activities to the ethical training of AI actors<sup>1</sup>. In that sense, it offers a valuable instrument to foster human agency and oversight over the whole AI life cycle [3], and reconnects the idea of responsible AI to the role of the acting subject [4].

The framework springs from classical philosophical traditions (for a synthesis see [5]), that emphasize the anthropological aspects of moral decisions. In this view, ethical judgment results from a process of discernment on what is “good” in a particular situation, facing conflicts with existing rules and seeking consultation with peers. The analysis of this process points out a reserve of ethical tools which are inherent within the human being and can play a meaningful role in shaping human decisions in various domains, including AI development and deployment.

Additionally, the framework offers a conceptual map for the definition of a set of “moral exercises” for the moral training of human actors involved in the decision-making process.

This work aligns with previous research highlighting the failures of abstracting AI systems from their social contexts, the so-called abstraction traps [6], and the tempting prospect of solving social problems through technical means[7] or a purely empirical perspective [8]. In particular, it connects to critical revisions of AI and tech ethics calling into question the narrow focus on procedures and design choices lacking substantive force of reform [9, 10, 11]. Our contribution is philosophical and practical. On the philosophical side, our framework recasts ethics in broader terms and rediscovers the element of personal commitment and intersubjectivity which is inherent in ethical reasoning and deliberation. We believe that this way of thinking can foster a more proactive form of responsibility that goes beyond legal duties set up by established norms. On the practical side, the way forward suggested by our framework solicits greater engagement in AI ethics activities and encourages the exercise of civic virtues breaking the barriers of domain-specific expertise or roles and pointing to common conditions (e.g. humanity and citizenship).

## 2. Human Oversight of AI systems

Human oversight is among top priorities in AI ethics guidelines spread worldwide [12] and intersects important principles such as transparency, justice and non-maleficence [13]. In the European Ethics Guidelines for Trustworthy AI, it aims at ensuring that the decisions aided by algorithms align with ethical principles and do not lead to harmful or undesirable outcomes [3]. The recent European legislation (AI Act) requires human oversight to prevent or minimize the risks to safety and fundamental rights posed by high-risk AI systems [14]. This requirement becomes even more urgent when decisions are fully automated and produce legal or significant effects on individuals and groups subjected to algorithmic decisions.

From a practical point of view, human oversight is primarily associated to the presence of a human decision-maker reviewing and validating the algorithmic outcome. The extent of human

---

<sup>1</sup>We consider AI actors as “those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI” [2].

oversight can vary, ranging from intervention in every decision cycle of the system (Human-in-the-loop) to the monitoring of the overall activity of the system, including its economic and societal impact (Human-in-command). For high-risk AI systems, the AI Act requires that humans in charge of supervision meet specific conditions (see article 14 [14]). For instance, they should have sufficient knowledge about the relevant capacities and limitations of the system, be aware of the possible tendency of overlying in the system's output (automation bias) and be able to override or reverse the output generated. Unfortunately, the proven limitations of human capabilities in assessing the quality of algorithmic output raised concerns upon the real expectations for human oversight and effective accountability [15, 16].

To overcome the possible flaws in human oversight, organizations introducing algorithmic decision making systems could pursue multiple strategies. They could improve the technical training of human operators or seek better integration with organizations' policies and procedures ensuring, e.g., transparency. On the other hand, the implementation of human oversight extends far beyond the introduction of a human overseer and involves various activities such as the set up of redress measures and communication channels. In addition, the undertaking of human oversight presupposes the fulfilment of other important tasks such as risk management and impact assessment. When it comes to the public sector, [16] propose to turn towards an institutional oversight approach based on evidence-based justifications and democratic review.

Here, we focus on the ethical dimensions involved in the oversight process. The (training) activities derived from this analysis target moral dispositions rather than technical skills or knowledge. They are not meant to overcome the challenges associated to poor cognitive capabilities of human operators. They aim at fostering a pedagogy of ethical practices [17] and support the growth of AI actors' responsibility over the life cycle of an AI-aided decision system. Our approach to human oversight is broad and covers all relevant choices influencing the algorithm's behaviour and the outcome, not only the final output.

### 3. A framework for Human Oversight

This framework offers a conceptual map that highlights different types of questions and ways of thinking when making ethical decisions. Its main purpose is to offer the conceptual coordinates to the moral exercises and mark differences with other approaches put forward in the field of AI ethics. A distinct character of this framework is to bring attention towards activities and dispositions shaping human behaviour. While a large contribution in the field of AI ethics consists of techniques or methodologies aimed at generating an external outcome (e.g. fairer or more intelligible outputs), our framework tries to intervene on internal resources stimulating changes in the acting agent.

#### 3.1. Structure of the framework

Our framework, depicted in the upper part of Figure 1, rests on two focal points which refer to the sphere of values and the sphere of actions, respectively.

**Values.** The first element encompasses questions addressing the ends of human life and the moral coordinates for human actions. The sphere of values sets the stage for moral actions by establishing a kind of "pre-normative" ethics placed on top of rights and duties set up by a

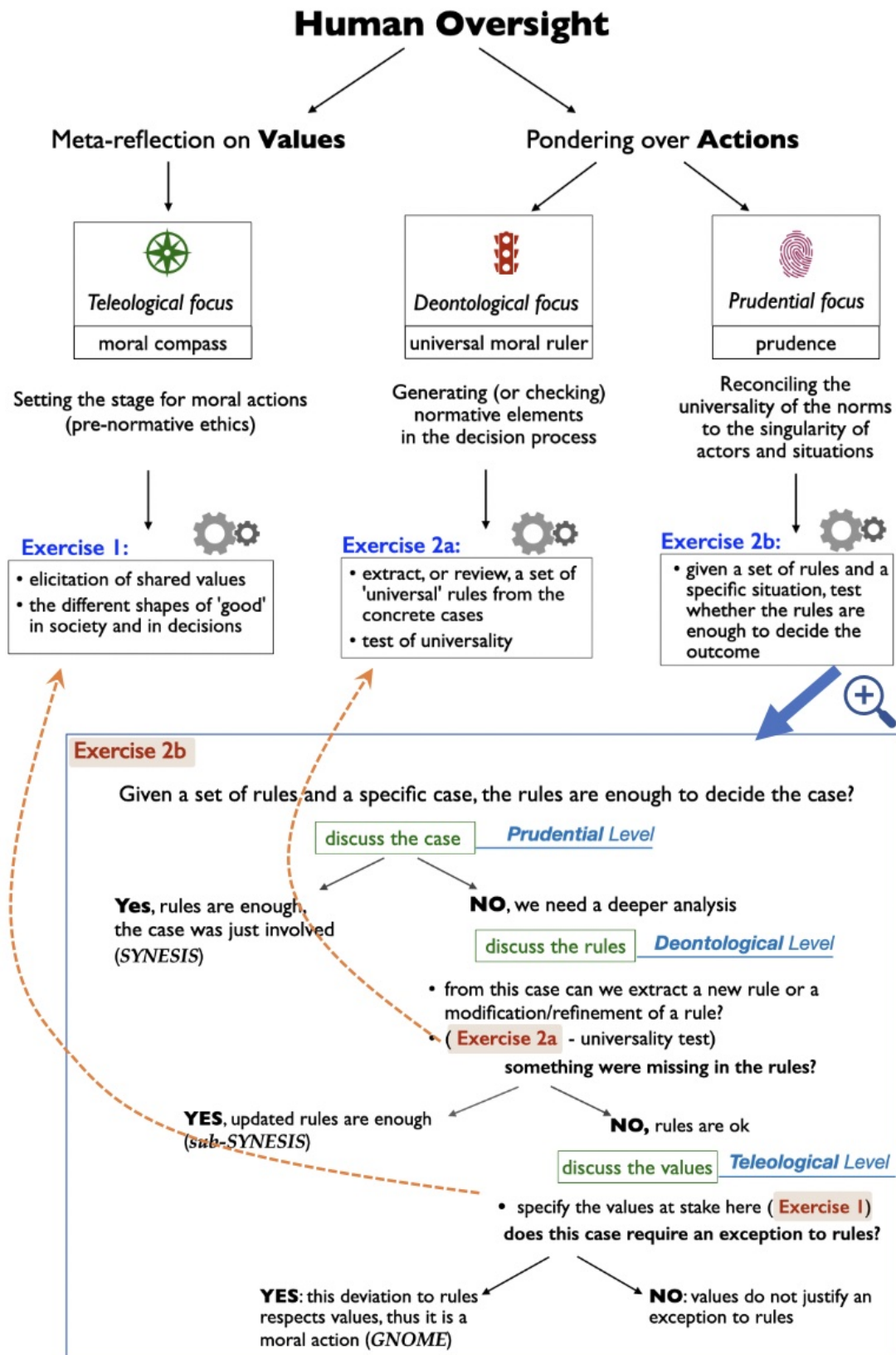


Figure 1: Ethical framework (upper part) and moral exercises (lower part)

social contract. In philosophy, it relates to Aristotle's theory (teleology) and stresses the aim of ethics: "a good life lived with and for others in just institutions." [18]. From a practical point of view, with this element we refer to a broad reflection on the ends (telos) of human decision and actions including consideration of the self, the others and common ways of life. Note that by "telos" we mean not only the ultimate purpose of human actions but also what one would consider morally important and worthy of pursuit, i.e. values. Indirectly, a reflection on values would also imply a scrutiny of disvalues, i.e. the possibility of harm and wrongdoing, to some extent. For this reason, this sphere could also be understood as a reflection on the various manifestations of good and evil in human experience.

The key tool here is the *moral compass*, a sort of internal sense or guide that help humans to recognize both "evil" and "good" without being irreparably confused nor blocked by the fact that they take different forms. Consider, for example, efforts to embed values into the design process of a technical artifact [19]. This may require in-depth sessions of stakeholder consultation in which people express their expectations, but also values and fears [20]. A case of participatory design for algorithmic decision-making in kidney transplant showed how people can share different meanings of health and fairness, address tensions and difficult trade-offs [21].

**Actions.** The second element deals with questions pondering concrete actions over determinate circumstances. It comprises a variety of questions, such as what is the right thing to do in this situation? Is this decision morally acceptable? We distinguish two classes of problems that subdivide ethical reasoning in two branches.

- The first problem is the definition of norms given a specific situation (e.g. communicating an adverse diagnosis to a patient) to assess the moral sustainability of actions with respect to the self and the others. This branch highlights the normative elements of a decision process and clearly reminds of Kant's moral imperatives (deontology). It requires to control the decision process from the viewpoint of universality and grounds moral obligations on the value of human dignity. This part of the framework puts forward a *universal moral ruler* as an ethical tool protecting the moral judgement from arbitrariness, violence and injustice. A genuine result of this ruling activity is the definition of codes of ethics and conduct in high-stake domains, such as medicine. Examples of code of ethics for computing professionals exist and spread across different associations [22, 23].
- The second issue is the determination of good and wrongdoing in specific situations and a given set of rules (e.g. communicating an adverse diagnosis to a remarkably fragile patient). In this case, humans have to reconcile the universality of the norms (the previous dimension) to the singularity of actors and situations. To overcome the limits of a purely deductive approach, this class of questions interrogates practical wisdom, also known as *prudence*, expressing human capacity to apply abstract rules in contingent situations. An important remark is that prudence builds upon a common inquiry which allows to collect and ponder different points of view - a character well reflected in Thomas Aquinas' concept of counsel (a "conference held between several" [24]). In the medical example, even if there is a general rule of transparency for doctors, the fragility of the specific patient could suggest that omitting some details of the diagnosis would improve the agency of the patient and its ability to recover. Consultation with other doctors and family members would also help address tensions and decide how values can be best

enacted in this specific circumstance.

### 3.2. Moral exercises

The conceptual framework outlined above suggests key ethical dynamics in decision-making processes. This conceptual description has a concrete counterpart in terms of activities that could be further elaborated into resources for the ethical training of human decision-makers. We call them *moral exercises* since they can inspire structured activities to help human actors master ethical skills, corresponding, in our case, to the tools suggested above (*the moral compass, the universal moral ruler, and prudence*). We identify an exercise for each component of the framework and summarize them in the lower part of Figure 1. Even if the exercises can be done independently, Figure 1 highlights their interconnected nature, mirroring the relationship between the different branches of the framework. The moral exercises were experimented in different social contexts that involve critical decision-making, such as social services and criminal justice [25]. In the following we illustrate the exercises and outline what they may look like in a medical domain.

**Exercise 1** seeks to identify a common set of shared values or potential harms in a context of teamwork. This activity starts with individual, subjective consultations and progresses to group discussions, often resulting in unanimity[25]. Note that the approach recommended by this exercise differs from the aggregation of ethical preferences facilitated by crowdsourcing platforms [26]. Instead, it aims to establish a consensus and foster a shared understanding of core values through dialogue leaving room for some disagreement.

Typically, such activities occur within small groups of peers. For example, in a medical setting, a team of medical professionals may want to establish a common ethical basis for making decisions in the hospital. They may converge on values such as loyalty towards patients, transparency, truthfulness, scientific accuracy, respect, active listening, attentiveness, and benevolence. Within the AI life cycle, this exercise may apply to the early developmental stages, for example when an organization has to define the key set of values guiding the whole AI project.

**Exercise 2a** focuses on the deontological level, aiming to derive one or more general rules based on a specific situation's perspective. This activity entails a group of individuals analyzing the potential universal applicability of a moral decision or action. In this exercise, the participants identify a course of action for a specific case and question whether the proposed solution would consistently align with shared values if adopted as a universal rule (*test of universalisation*).

For a concrete example of this exercise, consider a medical case with a 90% unfavourable diagnosis (e.g. processed by IBM Watson). The universalization test in this context involves determining which 'rule' (or Kantian 'maxim') becomes apparent, such as "tell the truth." Another rule could be "disclose that the diagnosis originates from an AI tool," but the discussion prompted by the analysis of an unfavorable diagnosis might lead to a refinement of the rule as follows: "disclose the algorithmic source of the diagnosis only if the doctor fully agrees with the automated assessment." With regard to the development of an AI system, this exercise would be useful in reviewing the model's performance in relation to the risks of unfairness. In this context, the AI actors fulfill various tasks that encompass identifying vulnerable groups, selecting fairness metrics, and evaluating trade-offs between accuracy and fairness.



We finally remark that Exercise 2a may reach different conclusions starting from the same initial conditions. This is perfectly valid, since there is not a single set of rules that is generally correct. Indeed, the goal of the exercise is not to define a universal procedure to test the ethics of decisions, but to engage subjects in a shared discernment process and foster a proactive form of responsibility.

**Exercise 2b** puts into practice the “prudential approach” which is distinctive of the classical philosophical tradition. Even with a moral guideline (possibly developed through Exercise 2a), moral judgment does not solely rely on logical deduction; it requires the use of practical wisdom. Consequently, Exercise 2b aims to cultivate ‘prudence,’ which enables one to transcend the limitations of rigid rules when their strict adherence might prove detrimental to the greater good in certain circumstances. Figure 1 illustrates this exercise, which revolves around discussing a conflicting case that necessitates consultation of various perspectives and careful consideration of its uniqueness and contingency. The result can be either the attainment of a favorable resolution in alignment with the provided moral coordinates (resembling Thomas Aquinas’ concept of *synesis* [24]), or the recognition that it might be necessary to depart from the initial norms in the pursuit of the greater good (akin to the concept of *gnome* [24]).

This exercise serves as a stress test for the moral ruler established as a deontological foundation, which is why it is closely related to Exercise 2a. The most challenging scenario emerges when one’s internal moral compass suggests that rigidly applying rules in a given situation could lead to harm. This is where prudence comes into play, as it is the human virtue that enables us to delve deeper and comprehend underlying values and meanings. Since straying from established rules can carry risks, the exercise highlights the importance of the teleological foundation as a safety net. In such cases, it requires revisiting the shared values determined in Exercise 1. Therefore, a final decision that deviates from the rules is only accepted if it is well-justified in terms of the underlying accepted values.

An illustrative application of this exercise is in the context of delivering an unfavorable diagnosis to a vulnerable patient. Specifically, the discussion can take into account the unique characteristics of the individual patient. This scrutiny might reveal that a subset of users had been overlooked in the formulation of general rules (e.g., individuals with Alzheimer’s disease or those with concurrent conditions), prompting a return to Exercise 2a. Conversely, if the existing rules prove sufficient, the team revisits the values at stake in the individual’s best interest, subsequently either confirming or rejecting the decision based on these rules. In both scenarios, the discussion bolsters awareness of the moral implications.

## 4. Discussion and concluding remarks

Our framework illuminates the ethical aspects of the oversight process, encompassing both values and actions, while shifting the focus from external factors to those within the individual. This highlights ethical dynamics and tools that can inspire human decisions and actions. In particular, our framework suggests the idea of moral exercises to support AI actors face important questions about the ends of an algorithmic-decision process or the exceptions that may result from the processing of certain input.

In this work, we have provided a broad overview, but our intention is to delve deeper into what moral exercises may entail in the oversight of an AI-assisted decision system. We have

some intuition about their potential utility in enhancing ethical dispositions in the oversight processes. We believe that structured activities could be elaborated around tasks comprising the developmental processes of an AI system. Current proposals of governance framework, such as [27], may provide useful insights to design scenarios for moral exercises, e.g. suggesting roles and AI-related tasks at different developmental stages. To this aim, we intend to expand upon and assess more comprehensive exercise proposals derived from pilot experiences in AI projects.

An IEEE report listening to engineers pointed out the need to allocate time for reflection and discussion enabling engineers' engagement and participation [28]. We believe that structuring activities in the form of Exercise 1 may help AI actors share their views on core values and elaborate a common understanding to support design choices. For example, similar elaboration might be beneficial when defining the annotation scheme for a data classification, a task that was considered a sense-making practice [29]. Awareness of different perspectives in ground-truthing [30] may call for a deeper understating of the values reflected in annotation or rating criteria and exercises on values could stimulate important insights in this regard. These examples are suggestive of a wide space for possible uses and further elaboration. To conclude, with this paper we aim to shift the focus from AI ethics techniques to AI actors, and engage researchers in a discussion on the extent to which our perspective can be explored and used.

## Acknowledgments

TS acknowledges financial support from the IRIS Academic Research Group (UK government, grant no. SCH-00001-3391). SC acknowledges the support of CINI National Laboratory Informatics & Society.

## References

- [1] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, *Science and engineering ethics* 26 (2020) 2141–2168.
- [2] Organisation for Economic Co operation, Development (OECD), Recommendation of the Council on Artificial Intelligence, 2019. URL: <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>, the 2019 OECD Ministerial Council Meeting.
- [3] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019. URL: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- [4] G. Gorgoni, R. Gianni, Responsibility, Technology, and Innovation, in: W. Reijers, A. Romele, M. Coeckelbergh (Eds.), *Interpreting Technology: Ricoeur on Questions Concerning Ethics and Philosophy of Technology*, Rowman & Littlefield, 2021, p. 171.
- [5] P. Ricoeur, *Éthique et morale*, *Revue de l'Institut Catholique de Paris* 34 (1990) 131–142.
- [6] A. Selbst, et al, Fairness and Abstraction in Sociotechnical Systems, in: *Proc. of the FAT\* conference*, 2019, pp. 59–68.



- [7] J. Powles, The seductive diversion of ‘solving’ bias in Artificial Intelligence, Medium (2018). URL: <https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53>.
- [8] T. Scantamburlo, Non-empirical problems in fair machine learning, *Ethics and Information Technology* 23 (2021) 703 – 712. doi:10.1007/s10676-021-09608-9, cited by: 3; All Open Access, Green Open Access, Hybrid Gold Open Access.
- [9] B. Green, The contestation of tech ethics: A sociotechnical approach to technology ethics in practice, *Journal of Social Computing* 2 (2021) 209–225.
- [10] T. Hagendorff, Blind spots in AI ethics, *AI and Ethics* 2 (2022) 851–867.
- [11] T. Scantamburlo, G. Grandi, A ‘little ethics’ for algorithmic decision-making, in: *CEUR Workshop Proceedings*, volume 3442, 2023. Cited by: 0.
- [12] T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines, *Minds and machines* 30 (2020) 99–120.
- [13] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, *Nature Machine Intelligence* 1 (2019) 389–399.
- [14] E. Commission, et al., Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, EC, COM 206 (2021) 21.
- [15] R. Koulu, Proceduralizing control and discretion: Human oversight in artificial intelligence policy, *Maastricht Journal of European and Comparative Law* 27 (2020) 720–735.
- [16] B. Green, The flaws of policies requiring human oversight of government algorithms, *Computer Law & Security Review* 45 (2022) 105681.
- [17] C. Huff, A. Furchert, Toward a pedagogy of ethical practice, *Communications of the ACM* 57 (2014) 25–27.
- [18] P. Ricoeur, *Oneself as another*, University of Chicago Press, 1992.
- [19] B. Friedman, Value-sensitive design, *interactions* 3 (1996) 16–23.
- [20] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, S. Mohamed, Power to the People? Opportunities and Challenges for Participatory AI, Equity and Access in Algorithms, Mechanisms, and Optimization (2022) 1–8.
- [21] D. G. Robinson, *Voices in the Code: A Story about People, Their Values, and the Algorithm They Made*, Russell Sage Foundation, 2022.
- [22] R. E. Anderson, ACM code of ethics and professional conduct, *Communications of the ACM* 35 (1992) 94–99.
- [23] E. W. Pugh, Creating the IEEE code of ethics, in: *2009 IEEE Conference on the History of Technical Societies*, IEEE, 2009, pp. 1–13.
- [24] T. Aquinas, *Summa theologiae*, i-ii; q. 14), 2023. URL: <https://aquinas.cc/la/en/~ST.I-II.Q14>.
- [25] G. Grandi, et al., *Fare giustizia. Un’indagine morale sul male, la pena e la riparazione*, Padova University Press, 2020.
- [26] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, *Nature* 563 (2018) 59–64.
- [27] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, Y. Wen, Capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act, Available at SSRN 4064091 (2022).
- [28] IEEE, *Addressing ethical dilemmas in AI: Listening to engineers*, 2020. URL: <https://>

[//standards.ieee.org/initiatives/artificial-intelligence-systems/ethical-dilemmas-ai-report.html](https://standards.ieee.org/initiatives/artificial-intelligence-systems/ethical-dilemmas-ai-report.html).

- [29] M. Miceli, M. Schuessler, T. Yang, Between subjectivity and imposition: Power dynamics in data annotation for computer vision, *Proceedings of the ACM on Human-Computer Interaction* 4 (2020) 1–25.
- [30] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, *arXiv preprint arXiv:2109.04270* (2021).