A 'Little Ethics' for Algorithmic Decision-Making

Teresa Scantamburlo[1], Giovanni Grandi[2]

[1]Ca' Foscari University of Venice and European Centre for Living Technology, via Torino 155, 30172 Venice, Italy
[2]University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy

Abstract
In this paper we present a preliminary framework aimed at navigating and motivating the ethical aspects of AI systems. Following Ricoeur's ethics we highlight distinct levels of analysis emphasising the need of personal commitment and intersubjectivity, and suggesting connection with existing AI ethics initiatives.

Keywords
Algorithmic decision-making, AI ethics, Paul Ricoeur,

1. Introduction

Ethics is playing a big role in Artificial Intelligence (AI) research. The growing awareness of the societal and environmental aspects of AI systems stimulated the development of concrete solutions to problems as diverse as algorithmic opacity and discrimination. In particular, huge efforts were directed to the development of ethical assessment or auditing methodologies (see [1, 2]) also in response to greater concerns about the spread of abstract ethical principles and a lack of practical guidance [3]. 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. Though operationalization recalls a distinctive character of applied ethics (i.e. that of being context-dependent and domain-specific), insisting on ethical solutions may leave for granted that dealing with the ethics of AI means, first and foremost, to identify and implement a set of ethical principles. In this paper we want to engage with more fundamental questions such as: What does it mean acting ethically in the context of AI? Or in other words, what does responsible behavior imply for AI innovation?

To address these questions we employ Paul Ricoeur's "little ethics" [4] which draws attention to fundamental elements of ethical decision-making and help us reconnect the notion of responsibility to the central role of the acting subject [5]. In particular, we distinguish three fundamental dimensions of ethical decision (teleological, deontological and prudential) and outline a preliminary framework to guide ethical reflection and actions regarding AI systems. The framework aims to stimulate engagement with questions that are marginally present in the AI ethics scholarship and deals with aspects of good life, shared societal values and tailored judgments. The framework is an instrument to exercise moral reasoning and deliberation in contexts of shared, common AI related efforts (e.g. groups involved in design processes, audits or ethics committee). It springs from key philosophical concepts which adapt to various domain, including AI, and touches upon questions regarding the ends of an AI system, the actors involved and the conflicts that may rise in specific contingencies (e.g. Whom is the purpose

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
teresa.scantamburlo@unive.it (T. Scantamburlo); giovanni.grandi@units.it (G. Grandi)

good for? Should the system be developed or used? Under which circumstances should the system be avoided or used differently?).

Here we focus on a particular class of AI systems, i.e. those which are used to support human-decision making, also known as prediction-based decision systems. Nevertheless, the framework may apply to other AI applications, stand-alone or embedded in more complex systems (e.g. automated vehicles), and focus on specific developmental stages of a system (e.g. design or evaluation). The framework is meant to address a broad spectrum of people to whom we will refer as the AI actors - i.e. "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI." [6]. To make our reasoning more concrete, we will consider an hypothetical scenario adapted from the case of UK's grading algorithm introduced in 2020 during the covid-19 pandemic [7].

2. Related works and motivations

This work aligns with previous research highlighting the failures of abstracting AI systems from their social contexts, the so-called abstraction traps [8], and the tempting prospect of solving social problems through technical means. 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].

While we acknowledge that there are multiple and diverse reasons behind this narrow conceptualization, our framework primarily reacts to the widespread, classical utilitarian assumption that treats morality as a formal calculus aimed at maximizing goods (the so called utility). This view is particularly attractive to computer science because it offers the conceptual basis to frame moral evaluation as a mathematical function which is neutral with respect to the content of subjects' preferences and centred on optimized trad-offs. [11]).

In contrast with the utilitarian perspective, Ricouer's synthesis recall philosophical traditions, such as those based on Aristotle and Kant, which shift the focus from the formality (e.g. how to maximize ends) to the content of moral action (e.g. which ends to choose). Also, these approaches stimulate an ethical discussion which goes beyond a purely economic and legal stance (where the logic of cost minimization and indemnity usually prevail) and invites reflection on the causes of harms and the actions that could prevent them.

Much of contemporary ethics rests on the assumption that values are inherently subjective and deliberating about common good is almost impossible to the point that aggregating individual preferences through large-scale mechanisms could appear an obvious solution for addressing moral AI-related dilemmas [12]. With our framework we aim to a different way of thinking which looks more positively to the elaboration of value judgments and the possibility of creating constructive dialogues among different, even contrasting, positions. [13]

Our contribution is philosophical and practical. On the philosophical side, our framework recasts ethics in broader terms and rediscover 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). In this sense, the framework can be understood as a meta-discourse that could help the discernment and the articulation of different moral instances involving personal aspirations, universal norms and the need of reconciliation.

3. A threefold ethical framework for AI-based decisions

Ricoeur's perspective has been influential in different fields of study. It was introduced in the review of health care practices [14] and, more recently, in the philosophy of technology to propose a new research program [15] and a narrative conceptualization of AI [16]. Here we recall Ricoeur's account of ethics which suggests an essential dynamics for decision and action. This dynamics flows through three interconnected dimensions or levels: it starts with the ethical aim, passes through the sieve of the norm and turns to practical wisdom (or prudence) for the concrete application [17]. This dynamics could be explored both as a descriptive and as a normative account, but this would require further elaboration which is out of the scope of this work. The dynamics may refer to a human process (e.g. an individual or collective decision) but also to an algorithmic process (e.g. an automated decision-making system). In this paper we are concerned with the first case, i.e. we will consider how the three dimensions of the framework interrogate humans when making a decision about a prediction-based decision system.

3.1. The dimensions of ethical decision-making

We now present the three dimensions by recalling key concerns and the role they play in the decision process. They could be understood as sequential steps but temporal aspects are not considered for the moment - the interaction among dimensions may involve back-and-forth interactions. The dimensions connect to existing AI ethics initiatives and methodologies, and can provide them with a broader horizon of meanings and sense.

Teleological. The first dimension sets the stage for ethical decision-making and recalls the aim of ethics, i.e. "a good life lived with and for others in just institutions." [18]. Built upon Aristotle's work, this dimension establishes a kind of "pre-normaitve" ethics placed on top of rights and duties set up by a social contract. It invites us to reflect on the end (*telos*) of a decision with respect to the self, the others and common customs. This dimension connect to the development of AI systems embracing a humanistic purpose (e.g. AI for social good projects or Human-centred AI research) and methodologies supporting proactive consideration of societal values into the design process [19].

Deontological. Based on Kant's moral imperatives, this dimension highlights the normative element of a decision process and raises awareness on two main points: the moral sustainability of one's action and the respect owe to others. It requires to control the decision process from the viewpoint of universality and grounds moral obligations on the value of human dignity. The dimension offers criteria of validity to the ethical intention (expressed in the first dimension) with a view to protect moral judgment from arbitrariness, violence and injustice. Various ethical principles proposed so far [20] fulfil precisely this task. Testing the sustainability of algorithmic decision is an active and critical exercise which seek and motivate the principles most relevant to achieve the ethical aim in a specific context.

Prudential. The third dimension tries to reconcile the universality of principles (stressed by

the second dimension) to the singularity of actors and situations. To overcome the limits of a purely principled approach, this dimension recalls the role of practical wisdom, also known as prudence, tasked with the application of abstract rules in contingent situations. An important remark is that prudence grows out of a common inquiry which allows to collect and ponder different points of view - Ricoeur evokes Thomas Aquinas' concept of counsel ("a conference held between several" [21]). This dimension sets the ground for participatory design practices aiming at increasing diversity and inclusion in the whole life cycle of the AI system [22].

3.2. A use case example

A possible use of the framework consists in guiding AI actors when they need to deliberate about an AI system (e.g. identifying and prioritizing issues). Consider, for example, the case of an ethics committee tasked with the review of an automated decision system for grading similar to [7]. Suppose that, after consulting technical experts and documentation, the committee has to deliberate whether recommending or not that application. We provide a tentative description of topics and issues solicited by the three dimensions.

The *teleological* dimension directs attention to the purpose of the system and to the purpose attached to students' assessment. Ethical issues of interest here include: to what extent these purposes relate to good life and common good, problems of misalignment between the purpose of the system and the purpose of assessment, the values that should be promoted, as well as the virtues that should be honoured, through grading (e.g. success, human flourishing, autonomy).

The *deontolgical* dimension suggests consideration of relevant rules and obligations in the context of education. Fairness and transparency are two obvious principles that would apply here. However, the dimension recalls a more radical stance than a scrutiny of applicable principles. For example, one may argue that automated grading would impinge human dignity making the assessment task impersonal separated from a student-teacher relationship. Further concerns would regard the long-term effect on students' mentality (more competitive behaviour, marginalization of less-performing people, etc), the impact of creating uniform standards missing special talents (e.g. one of the assumption in the UK's model was that schools tend to get the same kinds of students over time).

The *prudential* dimension points to the need of situated moral judgments which consider the specific context of action. Automated grading could be recommended in exceptional situations such as the covid-19 crisis. In any case, its adoption might be excluded in particular conditions as shown by the UK's controversy (e.g. when the number of students to be assessed is too small). Other important concerns relate to stakeholder consultations (e.g. who was consulted, how inputs were considered, etc) and the quality of stakeholders' engagement.

4. Conclusion

In this work we outline a preliminary framework inspired by Ricoeur's little ethics. The framework consists of three dimensions that cover distinct aspects of ethical decision-making. The dimensions can be used to guide ethical reflection on AI systems and be adapted depending on the task at hand (ethics review or design), the application's developmental stage, and the AI actors involved. The framework encourages the exercise of practical judgment and dialogue among AI actors going beyond formal procedures and risk management.

References

[1] I. D. Raji, et al, Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proc. of the FAT* 2020 conference, 2020, pp. 33–44.

[2] D. Peters, K. Vold, D. Robinson, R. A. Calvo, Responsible ai—two frameworks for ethical design practice, IEEE Transactions on Technology and Society 1 (2020) 34–47.

[3] 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.

[4] P. Ricoeur, Éthique et morale, Revue de l'Institut catholique de Paris 34 (1990) 131–142.

[5] 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.

[6] OECD, Recommendation of the council on artificial intelligence, 2019. URL: https://legalinstruments.oecd.org/en/instruments?mode=advanced&typeIds=2.

[7] E. F. Studio, Can an automated algorithm make human grading farier?, 2020. URL: https://www.edufuturesstudio.com/uk-exam-algorithm-game.

[8] A. Selbst, et al, Fairness and abstraction in sociotechnical systems, in: Proc. of the FAT* conference, 2019, pp. 59–68.

[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] J. Nida-Rümelin, N. Weidenfeld, Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age, Springer Nature, 2022.

[12] 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.

[13] S. M. J., Justice: What's the Right Thing to Do?, Farrar, Straus and Giroux, 2009.

[14] I. Ekman, Practising the ethics of person-centred care balancing ethical conviction and moral obligations, Nursing Philosophy 23 (2022) e12382.

[15] W. Reijers, A. Romele, M. Coeckelbergh, Interpreting technology: Ricoeur on questions concerning ethics and philosophy of technology, Rowman & Littlefield, 2021.

[16] M. Coeckelbergh, Time machines: Artificial intelligence, process, and narrative, Philosophy & Technology 34 (2021) 1623–1638.

[17] D. Pellauer, B. Dauenhauer, Paul Ricoeur, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2022.

[18] P. Ricoeur, Oneself as another, University of Chicago Press, 1992.

[19] E. Aizenberg, J. Van Den Hoven, Designing for human rights in ai, Big Data & Society 7 (2020) 2053951720949566.

[20] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399.

[21] T. Aquinas, Summa theologiae, i-ii; q. 14), 2023. URL: https://aquinas.cc/la/en/~ST.I-II.Q14.

[22] A. Birhane, et al, Power to the people? opportunities and challenges for participatory ai, Equity and Access in Algorithms, Mechanisms, and Optimization (2022) 1–8.