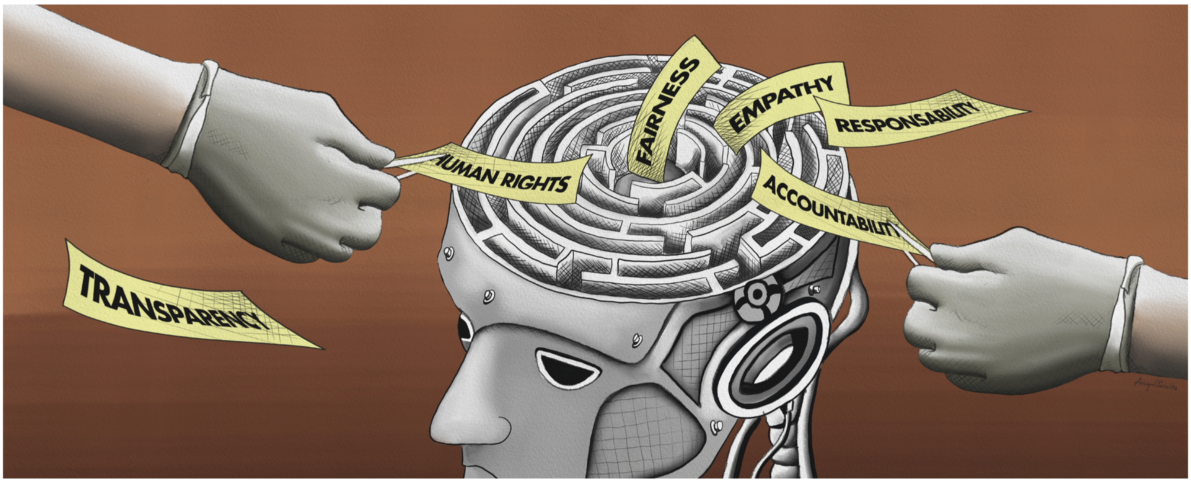
ETHICAL CHALLENGES

# Thinking About ‘Ethics’ in the Ethics of AI

Pak-Hang Wong, Judith Simon



Araya Peralta

A major international consultancy firm identified ‘AI ethicist’ as an essential position for companies to successfully implement artificial intelligence (AI) at the start of 2019. It declares that AI ethicists are needed to help companies navigate the ethical and social issues raised by the use of AI [1]. The view that AI is beneficial but nonetheless potentially harmful to individuals and society is widely shared by the industry, academia, governments, and civil society organizations. Accordingly and in order to realize its benefits while avoiding ethical pitfalls and harmful consequences, numerous initiatives have been established to a) examine the ethical, social, legal and political dimensions of AI and b) develop ethical guidelines and recommendations for design and implementation of AI [2] .

However, terminological issues sometimes hinder the sound examination of ethical issues of AI. The definitions of ‘intelligence’ and ‘artificial intelligence’ often remain elusive, and different understandings of these terms foreground different concerns. To avoid confusion and the risk of people talking past each other, any meaningful discussion of AI Ethics requires the explication of the definition of AI that is being employed as well as a specification of the type of AI being discussed. Regarding the definition, we refer to the European Commission High-Level Expert Group on Artificial Intelligence, which defines AI as “software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected [...] data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions" [3].

To provide specific guidance and recommendations, the ethical analysis of AI further needs to specify the *technology*, e.g. autonomous vehicles, recommender systems, etc., the *methods*, e.g. deep learning, reinforcement learning, etc., and the sector(s) of application, e.g. healthcare, finance, news, etc. In this article, we shall focus on the ethical issues related to *autonomous AI*, i.e. artificial agents, which can decide and act independently of human intervention, and we shall illustrate the ethical questions of autonomous AI with plenty of examples.

Consider first the case of autonomous vehicles (AVs). The possibility of accident scenarios involving AVs, in which they would unavoidably harm either the passengers or pedestrians, has forced researchers and developers to consider questions about the ethical acceptability of the decisions made by AVs, e.g. what decisions should AVs make in those scenarios, how can those decisions be justified, which values are reflected by AVs and their choices, etc [4].

> Hiring algorithms typically function by using the criteria they learned from a training dataset. Unfortunately, such training data can be biased, leading to potentially discriminatory models

Or, consider the case of hiring algorithms, which have been introduced to automate the process of recommending, shortlisting, and possibly even selecting job candidates. Hiring algorithms typically function by using the criteria they learned from a training dataset. Unfortunately, such training data can be biased, leading to potentially discriminatory models [5].

In order to ensure protection from discrimination, which is not only a human right, but also part of many countries' constitutions, we therefore have to make sure that such algorithms are at least non-discriminatory but ideally also *fair*. There are, however, different understandings of fairness: people disagree not only what fairness means, the adequate conception of fairness may also depend upon the context. Moreover, it has also been shown that different fairness metrics cannot be attained simultaneously [6]. This raises the question how values such as fairness should be conceived in which context and how they can be implemented.

One of the fundamental questions in the ethics of AI, therefore, can be formulated as a problem of value alignment: how can we build autonomous AI that is aligned with societally held values [7]. Virginia Dignum has characterized three dimensions of AI Ethics, namely "Ethics by Design", "Ethics in Design", and "Ethics for Design" [8], and they are useful in identifying two different responses to the value alignment problem. We shall structure the following discussion based on the three dimensions above and explore the two different directions to answer the value alignment problem in more detail.

## Building Ethical AI: Prospects and Limitations

Ethics by Design is “the technical/algorithmic integration of reasoning capabilities as part of the behavior of [autonomous AI]” [9]. This line of research is also known as ‘machine ethics’. The aspiration of machine ethics is to build artificial moral agents, which are artificial agents with ethical capacities and thus can make ethical decisions without human intervention [10]. Machine ethics thus answers the value alignment problem by building autonomous AI that by itself aligns with human values. To illustrate this perspective with the examples of AVs and hiring algorithms: researchers and developers would strive to create AVs that can reason about the ethically right decision and act accordingly in scenarios of unavoidable harm. Similarly, the hiring algorithms are supposed to make non-discriminatory decision without human intervention.

Wendell Wallach and Colin Allen classified three types of approaches to machine ethics in their seminal book *Moral machines* [11]. The three types of approaches are, respectively, (i) top-down approaches, (ii) bottom-up approach, and (iii) hybrid approaches that merge the top-down and bottom-up approach. In the simplest form, the top-down approach attempts to formalize and implement a specific ethical theory in autonomous AI, whereas the bottom-up approach aims to create autonomous AI that can learn from the environment or from a set of examples what is ethically right and wrong; finally, the hybrid approach combines techniques and strategies of both the top-down and bottom-up approach [12].

A These approaches, however, are subject to various *theoretical* and *technical* limitations. For instance, top-down approaches need to overcome the challenge to find and defend an uncontroversial ethical theory among *conflicting* philosophical traditions. Otherwise the ethical AI will risk being built on an *inadequate*, or even *false*, foundation. Bottom-up approaches, on the other hand, infer what is ethical from what is *popular*, or from what is *commonly held as* being ethical, in the environment or among examples. Yet such inferences do not ensure that autonomous AI acquire *genuine* ethical principles or rules because neither popularity nor being considered ethical offers an appropriate ethical *justification* [13]. Furthermore, there is the *technical* challenge of building an ethical AI that can effectively discern *ethically relevant* from *ethically irrelevant* information among the multitude of information available within a given context. This capacity would be required for the successful application of ethical principles in top-down approaches as well as for the successful acquisition of ethical principles in bottom-up approaches [14].

> Autonomous AI in general, and ethical AI in particular, may significantly undermine human autonomy because the decisions made by them for us or about us will be beyond our control

Besides the theoretical and technical challenges, several *ethical* criticisms have been leveled at building autonomous AI with ethical capacities. First, autonomous AI in general, and ethical AI in particular, may significantly undermine human autonomy because the decisions made by them *for us* or *about us* will be beyond our control, thereby reducing our independence from external influences [15]. Second, it remains unclear who or what should be responsible for wrongful decisions of autonomous AI, leading to concerns over their impacts on our moral responsibility practices [16]. Finally, researchers have argued that turning autonomous AI into moral agents or moral patients unnecessarily complicates our moral world by introducing in it unfamiliar things that are foreign to our moral understanding, thereby imposing an unnecessary ethical burden on human beings by requiring us to pay undue moral attention to autonomous AI [17].

## Machine Ethics, Truncated Ethics

Our review of the theoretical, technical, and ethical challenges to machine ethics does not intend to be exhaustive or conclusive, and these challenges could indeed be overcome in future research and development of autonomous AI. However, we think that these challenges do warrant a pause and reconsideration of the prospects of building ethical AI. In fact, we want to advance a more fundamental critique of machine ethics before exploring another path for answering the value alignment problem.

Recall the objective of machine ethics is to build an autonomous AI that can make ethical decisions and act ethically without human intervention. It zooms in on imbuing autonomous AI the capacities to make ethical decisions and perform ethical actions, which reflects a peculiar understanding of ‘ethics’ we take to problematize. More specifically, focusing *only* on capacities for ethical decision-making and action, machine ethics is susceptible to a *truncated* view of ethics that sees ethical decisions and actions as separable from their social and relational contexts. Philosopher and novelist Iris Murdoch, for example, has long ago argued that morality is not about “a series of overt choices which take place in a series of specifiable situations” [18], but about “self-reflection or complex attitudes to life which are continuously displayed and elaborated in overt and inward speech but are not separable temporally into situations” [19]. For Murdoch, what is ethical is inherently tied to a background of values. Therefore, it is essential, in thinking about ‘ethics’, to look *beyond* the capacities for ethical decision-making and action and the moments of ethical choice and action and *into* the background of values and the stories behind the choice and action. Similar arguments have been made to affirm the role of social and relational contexts in limiting ethical choices and shaping moral outcomes, and thus the importance to account for them in our ethical reflection [20].

Following this line of criticism, the emphasis on imbuing autonomous AI’s ethical capacities in machine ethics can be viewed as wrongheaded insofar as the emphasis overshadows the fact that ethical outcomes from autonomous AI are shaped by multiple, interconnected factors external to its ethical reasoning capacities and that there is an extended process of social and political negotiation on the criteria for rightness and wrongness underlining the eventual ethical decisions and actions made by autonomous AI. ‘The Moral Machine experiment' conducted by researchers at the MIT Media Lab is a case in point [21]. In the experiment, the MIT researchers attempt to crowdsource ethical decisions in different accident scenarios involving AVs, and the results are intended to inform the ethical design of AVs. What is missing, however, are the social, cultural, political backgrounds and personal stories involved in *real* accidents that accident scenarios in the experiment do not, and often cannot, properly describe [22]. In this respect, 'The Moral Machine' experiment is also based on a truncated view of ethics, which *only* considers the choice to be made in specific situations and neglect the background of values and contextual details that are essential for making ethical judgments.

> In thinking about 'ethics', it is essential to look beyond the capacities for ethical decision-making and action and the moments of ethical choice and action, and into the background of values and the stories behind the choice and action

Indeed, social and relational contexts matter to the ethical analysis of autonomous AI both *before* and *after* its implementation. For example, one can devise an impartial hiring algorithm, which assesses job candidates *only* on the basis of the qualities required by an opening. This impartial hiring algorithm could nonetheless remain discriminatory, and therefore ethically dubious, if the specific qualities required by the opening are inadvertently linked to race, gender, and social class. In this case, care must be taken not to reproduces the *pre-existing social bias* in the hiring algorithm. Moreover, even the best-intended technologies can bring serious adverse impacts to their (non-)users as bias and harm could *emerge* from the interaction between technology and the users and society [23]. Imagine an app which residents can use to report incidents, such as road damages to the local city council, which then uses an algorithm to sort and rank local problems based on those reports. If we assume that access to smartphones and thus to the app is unequally distributed, this may lead to underreporting of problems in areas with poorer residents. If not taken into account in the algorithmic sorting and ranking, this bias in the input data could then further increase inequalities between more and less affluent areas in the city [24].

The key lesson from the two examples is that having some ethical principles or rules inscribed in autonomous AI is insufficient to resolve the value alignment problem because the backgrounds and contexts *do* contribute to our overall judgment of what is ethical. We should remind ourselves that autonomous AI is *always* situated in some broader social and relational contexts, and so we cannot *only* focus on its *capacities* for moral decision-making and action. We need to consider not only *what* decisions and actions autonomous AI should produce, but also (i) *why* we—or, the society—think those decisions and actions are ethical, (ii) *how* we arrive at such views, and (iii) *whether* we are justified in thinking so. Accordingly, 'The Moral Machines' experiment is objectionable as it unjustifiably assumes that the most *intuitive* or *popular* response to the accident scenarios is the *ethical* response. Indeed, the reframing of questions gives us two advantages. First, we can now easily include other parties and factors *beyond* the autonomous AI in our ethical reflection. Second, it also makes explicit the possibility of (re-)negotiating which ethical principles or rules should be inscribed in autonomous AI (or even questioning the use of autonomous AI in a specific context altogether).

## A Distributed Ethics of AI

To be clear, we do not deny the need to examine the values embedded in technology and the importance to design and build technology with values that are aligned with human interests [25]. As the examples in this article show, autonomous AI can play a role in ethical decision-making and may lead to ethically relevant outcomes, so it is necessary to both examine the values embedded in it and to use shared societal values to guide its design and development. We do, however, want to question the aspiration of *delegating* ethical reasoning and judgment to machines, thereby stripping such reasoning and judgment from the social and relational contexts. A proper account of the ethics of AI should expand its scope of reflection and include other parties and factors that are relevant to the ethical decision-making and have contributed to the ethical outcomes of autonomous AI. To this end, it is essential for the ethics of AI to include various stakeholders, e.g. policy-makers, company leaders, designers, engineers, users, non-users, and the general public, in ethical reflection of autonomous AI. Indeed, only by doing so can we sufficiently address the questions: (i) *why* we think the decisions and outcomes of AI are ethical, (ii) *how* we arrive at such views, and (iii) *whether* we are justified in our judgements.

> the design and implementation of AI should take existing societal inequalities and injustices into consideration, account for them, and at best even aim at alleviating them through their design decisions

We shall call this expanded AI Ethics a *distributed ethics of AI*. The term ‘distributed’ aims to capture the fact that multiple parties and factors are relevant to and have contributed to the ethical outcomes of autonomous AI, and thus the responsibility for them are ‘distributed’ between the relevant and contributing parties and factors [26]. To use the examples of AVs and hiring algorithms: poor urban planning and road facilities should be legitimate concerns in the ethics of AVs, in the same way as existing social and cultural biases are valid considerations for ethical hiring algorithms. Hence, the design and implementation of AI should take *existing* societal inequalities and injustices into consideration, account for them, and at best even aim at alleviating them through their design decisions.

The distributed ethics of AI needs what Dignum has labeled “Ethics *in* Design”, i.e. “the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures” as well as “Ethics *for* Design”, i.e. “the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems" [27]. Ethical questions of autonomous AI cannot be solved by 'better' *individual(istic)* ethical capacities but only through *collectiveefforts*. To guide such collective efforts, *ethicalguidelines* offer useful means to stir value- und principle-based reflection in regards in autonomous AI and to effectively coordinate the efforts among different relevant and contributing parts [28].

## Conclusions: sobre la IA fiable de la UE

In April 2019, the High-Level Expert Group released the 'Ethics Guidelines for Trustworthy AI' which concretize the Europe's vision of AI. According to these Guidelines, Europe should research and develop *Trustworthy AI*, which is *lawful*, *ethical*, and *robust*.

There are two points in the Guidelines that deserve special mentioning in the present subject of discussion. First, it is interesting to note that the concerns for trust in the Guidelines are about "not only the technology's inherent properties, but also the qualities of the socio-technical systems involving AI applications […]. Striving towards Trustworthy AI hence concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle." In this respect, the vision of Trustworthy AI clearly matches with the distributed ethics of AI as previously described. Second, it is also interesting to note that the four ethical principles identified in the Guidelines are *mid-level principles*, i.e.

1. The principle of respect for human autonomy.
2. The principle of prevention of harm.
3. The principle of fairness.
4. The principle of explicability

The formulation of ethical principles based on *mid-level principles* is particularly illuminating, because mid-level principles *require* human interpretation and ordering in their application, and they are not intended to—and, indeed cannot—be implemented within autonomous AI. The need for interpretation and ordering also points to the social and relational contexts, where the resourcesfor interpretation and ordering lies.

While the European vision of Trustworthy AI and the Guidelines have a conceptually sound foundation, there a number of open problems with them. For instance, the use of mid-level principles in the Guidelines allows considerable room for interpretation, which, in turn, can be misused by malevolent actors to cherry-pick the interpretations and excuse themselves from their responsibility. This problem is further compounded by the Guidelines' emphasis on self-regulation, where politicians and companies can pay lip service to the European vision with *cheap* and *superficial* measures, such as propaganda and setting up symbolic advisory boards, without *substantively* addressing the negative impacts of AI. Hence, there are significant issues concerning the *actual* regulatory and institutional framework for AI Ethics and for realizing this European vision. Particularly, there is the need to create a clear framework to *fairly* distribute the benefits and risks of AI and the need to introduce 'hard' laws and regulations against the violation of basic ethical values and human rights.

Notwithstanding these problems, the Guidelines' focus *on humans* and *beyond technology* should be taken as an appropriate *normative* standpoint for the AI Ethics and the European vision. To end this article, we want to remind that the ethical questions about autonomous AI are distributed in nature, and that we—or, the society—should have a voice in their design and deployment.

## REFERENCES

1 — KPMG (2019) Top 5 AI hires companies need to succeed in 2019.

2 — AlgorithmWatch has compiled a list of ethical frameworks and guidelines available at: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

3 — European Commission High-Level Expert Group on Artificial Intelligence [AI HLEG] (2019) Ethics guidelines for trustworthy AI. European Commission.

4 — The type of accident scenarios is known as 'the trolley problem'. It is only one of the topics discussed in the ethics of autonomous vehicles, and we only use it as an example to illustrate one of the many ethical issues autonomous AI could raise. See:
• Lin, P. (2016) Why ethics matters for autonomous cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 69-85). Berlin: Springer.
• Keeling, G. (2019) Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*.

5 — Bogen, M. (2019) All the ways hiring algorithms can introduce bias. *Harvard Business Review*, May 6, 2019.

6 — See:
• Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016) On the (Im)possibility of fairness. arXiv:1609.07236.
• Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2): 153-163.
• Wong, P.-H. (2019) Democratizing algorithmic fairness. *Philosophy & Technology*.

7 — The AI alignment problem is first explicitly formulated by Stuart Russell in 2014, see: Peterson, M. (2019) The value alignment problem: a geometric approach. *Ethics and Information Technology* 21 (1): 19-28.

8 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): 1-3.

Íbid., p. 2

10 — See:
• Winfield, A., Michael, K., Pitt, J., & Evers, V. (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE* 107 (3): 509-517.
• Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
• Misselhorn, C. (2018) Artificial morality. concepts, issues and challenges. *Society* 55 (2): 161-169.

11 — Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

12 — Íbid., p. 79-81

13 — For a review of the difficulty of machine ethics, see: Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019) Motivations and risks of machine ethics. *Proceedings of the IEEE* 107 (3): 562-74.

14 — This is also known as the moral frame problem, see: Horgan, T., & Timmons, M. (2009) What does the frame problem tell us about moral normativity? *Ethical Theory and Moral Practice* 12 (1): 25-51.

15 — Danaher, J. (2018) Toward an ethics of AI assistants: an initial framework. *Philosophy & Technology* 31 (4): 629-653.

16 — Matthias, A. (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6 (3): 175-83.

17 — Bryson, J. J. (2018) Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* 20 (1): 15-26.

18 — Murdoch, I. (1956) Vision and choice in morality. *Proceedings of the Aristotelian Society, Supplementary* 30: 32-58. p. 34

19 — Íbid., p. 40

20 — Walker, M. U. (2007) *Moral Understandings: A Feminist Study in Ethics*. Oxford: Oxford University Press.

21 — Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018) The Moral Machine experiment. *Nature* 563: 59-64.

Jaques, A. E. (2019) Why the moral machine is a monster. Paper presented to We Robot 2019, University of Miami, April 11-13, 2019.

23 — Friedman, B., & Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems* 14 (3): 330-347.

24 — Simon J (2012) E-Democracy and Values in Design. *Proceedings of the XXV World Congress of IVR 2012*.

Simon, J. (2017) Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The Ethics of Technology Methods and Approaches* (pp. 219-235). London: Rowman & Littlefield.

26 — See:

- Floridi, L. (2013) Distributed morality in an information society. *Science and Engineering Ethics* 19 (3): 727-743.
- Simon, J. (2015) Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (Ed.), *The Onlife Manifesto* (pp. 145-159). Cham, Springer.

27 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): p. 2.

28 — Floridi, L. (2019) Establishing the rules for building trustworthy. *Nature Machine Intelligence* 1: 261-262.

**Pak-Hang Wong**

Pak-Hang Wong is a philosopher of technology, graduated in Philosophy and History at the Hong Kong University. He is part of the Research Group for Ethics in Information Technology at Department of Informatics, Universität Hamburg, where he examines social, ethical, and political issues of algorithms, big data, robotics, artificial intelligence and other emerging technologies. He received his doctorate in Philosophy from the University of Twente in 2012, and held academic positions in Oxford and Hong Kong prior to his current position. He is the co-editor of *Well-Being in Contemporary Society* (2015, Springer) and has published in various academic journals.

**Judith Simon**

Judith Simon holds a bachelor in Psychology from the Freie Universität Berlin and a PhD in Philosophy from Universität Wien. She is currently professor of Ethics of Information and Technology at Universität Hamburg. She is also an editorial board member in magazines like *Philosophy and Technology* (Springer) and *Big Data & Society* (Sage), and she is member of the german Ethics Committee. She has been a scholar visitor at Stanford University annd guest researcher in Trento, Ljubljana and at Institut de Recerca en Intel·ligència Artificial (CSIC-IIIA) in Barcelona. Her main research areas are epistemological and ethical problems related to IT, communication, computing, computer ethics and ICT user and designer ethics. She was awarded with the Herbert A. Simon prize from the *International Association for Computing and Philosophy* (IACAP).