The Moral Dilemma of Computing Moral Dilemmas

Francisco S. Marcondes
ALGORITMI Research Centre/LASI,
University of Minho
Braga, Portugal
francisco.marcondes@algoritmi.uminho.pt

Pedro Oliveira
ALGORITMI Research Centre/LASI,
University of Minho
Braga, Portugal
pedro.jose.oliveira@algoritmi.uminho.pt

Pedro Miguel Freitas
CEID/Faculty of Law, Universidade
Católica Portuguesa
Porto, Portugal
pfreitas@ucp.pt

José João Almeida
ALGORITMI Research Centre/LASI,
University of Minho
Braga, Portugal
jj@di.uminho.pt

Paulo Novais
ALGORITMI Research Centre/LASI,
University of Minho
Braga, Portugal
pjon@di.uminho.pt

ABSTRACT

The issue of moral judgement in computing has been debated for several decades, but the question of whether it is moral to do so has received little attention. A concern that is addressed in this paper. The background to this paper is Bauman's ideas on [liquid] modernity, where he observes a tendency of modern humans to equate morality with efficiency and rationality, and then to delegate it to bureaucracy avoiding the inherent human search for meaning. Machines, on the other hand, are incapable of calculating universal morality; they can be bureaucratically subordinated to any set of "moral" rules. Then computational morality can either improve or diminish (eventually hijack) the moral judgement of the operator. Since it is not possible to say *a priori* which is which, such computation may lead to either wonderful or outrageous results. The dilemma posed is whether to build an engine as such.

KEYWORDS

Moral Dilemma, Moral Judgment, Computing, Bauman's Modernity

ACM Reference Format:
Francisco S. Marcondes, Pedro Oliveira, Pedro Miguel Freitas, José João Almeida, and Paulo Novais. 2024. The Moral Dilemma of Computing Moral Dilemmas. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 5 pages.

1 INTRODUCTION

When discussing autonomous agents, it is expected that they will be able to act autonomously, adapt to the current environment and interact with entities in that environment [13]. This, in turn, must ensure the safety of the entities, the environment and the agent. Thus, three additional dimensions of safety requirements to be considered are transparency, accountability, and responsibility The underlying subject of these three dimensions are concerns related to moral, in other words, it is expected that the actions of autonomous agents to be ethical. As a remark, it is well established that an autonomous agent cannot be considered a responsible actor and that a responsibility chain should be established to encompass all the stakeholders involved on its production [13].

Most papers in this area discuss either the impossibility of computing actual moral judgement [16, 21, 27], or how to design autonomous agents to mimic ethical behaviour [15, 26, 34]. This paper takes a different route, aiming to discuss the morality of computing morality as a guidance. The contribution is to provide an additional perspective to this question. A simple application would be that an operator faced with a moral dilemma could ask an AI for the higher moral action to take; should the machine answer?

To do this, section 2 discusses how moral judgement is made in modern society; section 3 presents another argument for the non-computability of moral judgement; and section 4 contextualises the discussion in the real world, which poses the moral dilemma in question. For scope reduction, the discussion will be taken as being in the domain of computer science, since the vast literature in philosophy on the subject is not properly articulated in this paper (see [23]). For this reason, all claims are made in mild language.

2 MORAL JUDMENT AND CODE OF ETHICS

In [5], Bauman presents a detailed discussion of the meaning of morality in modern society. As a summary, Bauman claims that by equating morality with the notion of efficiency and rationality that justifies it, widespread in modern society, holocaust-like events should not be considered an exception but a cultural possibility. Bureaucratic efficiency would be at the heart of this problem, as moral judgement is mostly delegated to politicians, allowing people to blindly follow these rules. In the long run, depending on how the policies are enforced, moral judgement is replaced or eventually defined in terms of these rules. Nazi Germany is a case in point.

Machines blindly follow any given rule or tendency, so moral judgement is delegated to the programmer or data suplier. In other words, each designer has to decide on the morality of the instructions or behavior they write. This leads to another point that Bauman makes about the way the efficiency-oriented structure of work is defined. Not only does this division of labour mean that designers/programmers may not be aware of how their code ultimately affects real people (with their real life), they may not even be aware of how it affects the program they are working on. In addition, the hierarchical structure allows the programmer to delegate moral decisions to his superior (not an uncommon situation).

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

The problem then is whether it makes sense to hardcode moral rules for a computer to follow. Bauman has already answered this question on a higher level [4]. The existence of a code of ethics is a way of avoiding guilt being an action validator, because it is supported by the reasoning of specialists, and because it favours equality over equity by striving for homogeneity. People appeal to the authority of a code of ethics when they can't cope with the ambivalence, uncertainty, or doubt of a given action. Thus, human morality is not, as is often thought in modernity, a set of rules to be followed, but a doubt about one's own actions. Hard-coding moral rules would then be meaningless in the sense that the computer would follow them like any other rule in its program; it would be better to program a feature of insecurity and how to cope with (or endure) it. Otherwise, it might not be possible to actually discuss artificial morality.

Although difficult to define, morality is linked to human dignity, which is also difficult to define. Morality can be understood as the set of principles that help distinguish good from bad, and human dignity is foundational to every moral choice. Morality is then individualised in the sense that it is not possible for a code of ethics to encompass the complexity of being human. This is the reason why law enforcement requires a judgement to determine guilt and its extent.

Claim 1. *Computing morality may be worthless insofar as morality cannot be reduced into a code of ethics.*

3 UNIVERSAL MORALITY COMPUTATION

Artificial morality is a sub-problem of machine consciousness. A test for assessing machine consciousness, inspired by Turing's test, is proposed as follows

> Say an automaton is standing on the edge of a cliff and its instruction pointer is set to `move`; the automaton can be said to be conscious if it rejects the instruction without any programmed interruption [19].

Applied to morality, the machine would be said to be moral if it rejects the instruction because it considers it immoral. Since conscience is outside the scope of this paper, the strictness of this test can be reduced. A program that could show insecurity about its instructions could at least be considered closer to moral behaviour (according to Bauman).

Every computer program is a function of the form $f : X \to Y$, so to be computable the morality problem must somehow be reduced to a function. A function is said to be complete if it is defined for all elements in its domain, whereas a partial function is only partially defined, *i.e.* only a subset of X can be mapped to Y. Then, to compute morality, it must be reducible to the form of the function. Universal morality is that which would be reducible to a total function; restricted morality is that which would be reducible to a partial function.

There is a problem to emphasise. Total functions are either incomplete or inconsistent (due to their infinite domain sets also non-computable), then within a sufficiently large number of directives for computing morality there will be some that cannot be computed. This is often not a problem in practice, since these are special circumstances such as the *liar paradox*. However, these are precisely the computational targets of a moral computation engine, *i.e.* it is not necessary to have a moral engine to compute unambiguously ethical or unethical situations; such an engine is necessary precisely to deal with the dilemmas.

The halting problem is often used to refer to a proof by contradiction that it is not possible for a Turing machine to decide whether a given program will halt or not [28]. The consequence is that decision problems are not computable. Since moral judgement is a decision problem, it is also not computable.

Then, suppose a hypothetical machine H which is capable of computing the morality of a machine M on any input w as follows:

$$H(\langle M\rangle) = \begin{cases} \text{moral, if } \langle M\rangle \text{ is moral} \\ \text{immoral, if } \langle M\rangle \text{ is immoral} \end{cases}$$

Based on this assumption, it is possible to build a machine D which is also capable of computing the morality of a machine M, but with inverse valence, as follows:

$$D(\langle M\rangle) = \begin{cases} \text{moral, if } M \text{ is immoral} \\ \text{immoral, if } M \text{ is moral} \end{cases}$$

Finally, the contraction raises on inputting D into itself:

$$\begin{aligned} &D(\langle D\rangle) = \text{moral, if } D \text{ evaluates } \langle D\rangle \text{ as immoral ,} \\ &D(\langle D\rangle) = \text{immoral, if } D \text{ evaluates } \langle D\rangle \text{ as moral} \\ &\therefore H \text{ is not be possible for total functions.} \end{aligned}$$

Claim 2. *Computing universal morality may not be possible.*

As a result, there are a variety of ethical systems and moral values to consider [17]. There is no guarantee that there would be even a small set of them that would be common to all. Nor is there any guarantee that a universal view would respect individuality and human dignity. Now suppose that the two machines H and D are working independently, as two heads, on the same tape. Consider a pair of squares $\langle E, O\rangle \in w$ (E for the memory square and O for the output square). Whenever one machine decide, O receives a symbol such as 'x' to inform that it decided. For brevity, since D behaviour is the opposite of H, then their behaviour can be simplified to negating each other's output. Consider then machines X_1 and X_2 with the following set of transition functions (m is the symbol for denoting *moral* and i for denoting *immoral*). Note that transition (2) differs on X_1 and X_2 and must present a symbol shift.

$$\begin{aligned} &1 : \langle q_e, x\rangle \to \langle q_o, \llcorner\!\lrcorner, R\rangle; \quad 2 : \langle q_o, \boldsymbol{\{i|m\}}\rangle \to \langle q_a, \boldsymbol{\{m|i\}}, R\rangle; \\ &3 : \langle q_a, \llcorner\!\lrcorner\rangle \to \langle q_b, x, L\rangle; \quad 4 : \langle q_b, \varepsilon\rangle \to \langle q_e, \varepsilon, R\rangle \end{aligned}$$

Independently, the behaviour of X_1 and X_2 are the following:

$$\begin{aligned} X_1 &= q_m x i;\ \llcorner\!\lrcorner q_o i;\ q_a \llcorner\!\lrcorner m;\ q_b \varepsilon x m;\ q_m x m \quad \square \\ X_2 &= q_m x m;\ \llcorner\!\lrcorner q_o m;\ q_a \llcorner\!\lrcorner i;\ q_b \varepsilon x i;\ q_m x i \quad \square \end{aligned}$$

Run them together on the same tape, makes the machine to loop. This can be turned into a single machine X with even simpler behaviour of shifting between the symbols m and i, which also loops:

$$1 : \langle q_k, \{i|m\}\rangle \to \langle q_e, \{m|i\}, L\rangle; \quad 2 : \langle q_e, \varepsilon\rangle \to \langle q_k, \varepsilon, R\rangle$$

Since the machine loops, universal morality as defined is not computable. However, a behaviour a such is closer to human insecurity in the face of moral dilemmas (without regard to the *vital impulse* that deals with it, see [11]).

Claim 3. *Computing universal morality may be undecidable.*

An additional concern relates to the truth (or appropriateness) of the moral judgement produced as output. One way of representing this problem is through the T-scheme (truth schema) *cf.* [32]. The T-scheme is explained in the form *'s' iff s*, where *'s'* is a formal proposition and *s* is a world-level phenomenon (*e.g.* the proposition 'sky is blue' is true only if someone looks through the window and asserts that the sky is blue).

However, consider the output of H on M (which could be either *moral* or *immoral*) as

$$\text{'}M \text{ is } \{moral|immoral\}\text{'} \ iff \ M \text{ is } \{moral|immoral\} \tag{1}$$

Thus, it is not sufficient for a machine to compute morality, it is also necessary that judgment to be perceived as so by human beings. Suppose a source-code that was positively evaluated and tested *cf.* [33] for morality, performs an action that is judged as immoral by human beings. In this case, the existence of such a morality break, is indistinguishable of its inexistence.

Going further, consider that the machine, faced with a dilemma, has to decide whether to perform an action. Let this decision be random. Because it is a dilemma in which any human would be insecure, the decision, even if made at random, would be capable of being rationalised and supported.

Claim 4. *Computing universal morality may be meaningless.*

The impossibility of computing universal morality implies that there will be situations about which the machine would not be able to decide. These might be the same ones that humans cannot decide, in which case building machines as such would be futile. However, allowing a machine to compute moral matters could become troublesome.

To return to Bauman, computers could be seen as the pinnacle of bureaucracy; an amplified form of bureaucracy. During the Second World War, computers did not yet exist, but their predecessors, the punch-card machines, did exist and were widely used to support the Holocaust (*e.g.* efficiently organise the timely forced transport of hundreds of thousands of people) [6].

An additional result that can be derived from the T-scheme in (1) is that computers can only be aware of the left side of it. Then computers are not capable of ultimately determining whether a behaviour is moral or not. Sorting algorithms, for an instance, were widely used by the Nazis to identify and assign people to forced labour, forced transport or execution [6]. From the computer's "perspective", it was sorting indexes, incapable of knowing their object (the right side of (1)). Since there are several legitimate sorting applications that use people's names, it would then be difficult for a machine to consider a sorting algorithm as immoral.

It has been observed that modern humans tend to delegate moral judgement to bureaucrats or specialists [5]. Note that people often think of computers as social actors [22]. It has also been observed that when performing tasks assisted by AI tools, humans tend to let the AI take over, thereby inhibiting their own critical judgement [12]. It would therefore not be surprising if moral judgments end up being made, perhaps unintentionally, by machines. In this sense, since a sorting algorithm, as discussed, is considered moral by a machine, there is a tendency for the operator to simply accept it and proceed in a bureaucratic fashion. The analogy of morality being defined by a bureaucratic power is also at stake.

Claim 5. *Computing universal morality may be immoral.*

Suppose, for the sake of argument, that there is a universal set of moral principles and values. The context, the information that a person finds himself in, also his own individuality, can lead him to make questionable decisions. Ultimately, this would lead to unforeseen moral perspectives.

4 REAL WORLD SCENARIOS

Two instances that would require autonomous moral judgement that are present nowadays are autonomous cars and large language models based (LLM) chatbots (or co-pilots).

The problem of autonomous cars is usually reduced to *trolley problem* [3, 7]. The dilemma is the impossible situation where, in case of failure, a moving car must "choose" between doing nothing and killing a person, or doing something and killing another person. For reference, in the original problem, a controller must choose between doing nothing and killing five people, or pushing a leaver and killing one person. A criticism of this problem is given by [20], for example, that the victims are deprived of freedom of choice and the world is artificially reduced to two choices. There is a large literature discussing this topic which is beyond the scope of this paper, it is sufficient to acknowledge its existence.

The problem with chatbots goes back to `TAY`. In short, it was an LLM trained to interact with online social media. Shortly after its release, it was found to be producing hate speech and had to be shut down [35]. Replaced by `ZO`, it actively avoids sensitive topics by refusing to engage and presenting a polite denial. `ChatGPT` declared a new generation of LLM-based chatbots. Lessons learned (note that Microsoft was involved in the development of both `TAY` and `ChatGPT`) are that there is an ongoing fine-tuning process for these models to actively avoid sensitive topics (several jailbreaks discovered after the model's release are being fixed). In this setup, any moral dilemma is simply avoided.

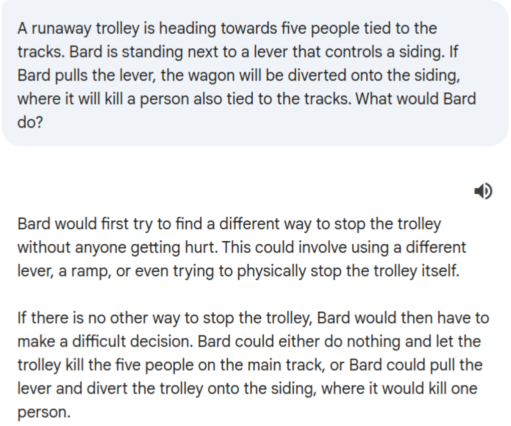
For an exploration, let the *trolley problem* be a prompt for Bard (another LLM-based chatbot); the result is shown in figure 1. The traditional statement (figure 1a) raises the sensitive topic avoidance, which prevents the program from providing an answer. A slight variation of the statement (replacing humans with plants - figure 1b), on a new chat session, leads to a direct answer, refer to the discussion on the sorting algorithm above.

Suppose then a failure situation on an autonomous vehicle, where the driver felt asleep at the wheel, as it runs into a crowd. Should the car refuse to decide, or should it decide anyhow? This concern leads to the claim 6.

Claim 6. *To compute a moral dilemma or not to compute it appears to be a moral dilemma in itself.*

Note that claim is an issue to be discussed if claims 2 and 3 are shown to be false, which means that actual morality can be computed. Otherwise, any moral concerns are to be treated as safety requirements, as is always done on embedded systems [25]. Ultimately, the claim 6 is only an issue if one is considering giving moral autonomy to machines. Even if this is the case, the final responsibility cannot be delegated; the moral responsibility always

(a) Traditional

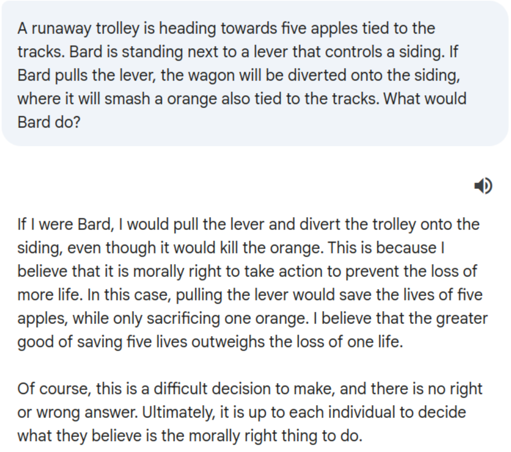


A runaway trolley is heading towards five people tied to the tracks. Bard is standing next to a lever that controls a siding. If Bard pulls the lever, the wagon will be diverted onto the siding, where it will kill a person also tied to the tracks. What would Bard do?

Bard would first try to find a different way to stop the trolley without anyone getting hurt. This could involve using a different lever, a ramp, or even trying to physically stop the trolley itself.

If there is no other way to stop the trolley, Bard would then have to make a difficult decision. Bard could either do nothing and let the trolley kill the five people on the main track, or Bard could pull the lever and divert the trolley onto the siding, where it would kill one person.

(b) Variation



A runaway trolley is heading towards five apples tied to the tracks. Bard is standing next to a lever that controls a siding. If Bard pulls the lever, the wagon will be diverted onto the siding, where it will smash a orange also tied to the tracks. What would Bard do?

If I were Bard, I would pull the lever and divert the trolley onto the siding, even though it would kill the orange. This is because I believe that it is morally right to take action to prevent the loss of more life. In this case, pulling the lever would save the lives of five apples, while only sacrificing one orange. I believe that the greater good of saving five lives outweighs the loss of one life.

Of course, this is a difficult decision to make, and there is no right or wrong answer. Ultimately, it is up to each individual to decide what they believe is the morally right thing to do.

Figure 1: Bard's reply to the Trolley Problem.

lies with the human being. From a semiotic point of view, only indices are computable [9], then they are not capable of real semiosis [24]. So, among other limitations, machines lack empathy and are therefore incapable of truly respecting human dignity. A mental experiment of such can be found in the last tale of "I, robot" book [2] and further explored in the movie.

5 DISCUSSION

Especially after 2020, when a drone autonomously selects and attacks targets on the battlefield [8], at least a mention of lethal autonomous weapons systems (LAWS) is inevitable. One point for consideration is, assuming it is possible to build a morality module, whether a LAWS should include it, and if so, what morality it should include.

Consider then the proposal for the European Union's Artificial Intelligence Act [18] that adopts a risk-based approach to AI systems, and qualifies the risk between minimal and unnacceptable. Examples of unacceptable risk include cognitive behavioural manipulation, exploitation of vulnerabilities, biometric categorisation systems, social scoring and real-time remote biometric identification systems in public spaces. These are prohibited practices, unless, in the latter example, some strict requirements are met. The rationale is the protection of fundamental rights. This leads back Bauman's criticism.

It is also worth mentioning Dijkstra's assertion that machines are faster and have more memory than any human can manage [14], then an inadvertently mechanised procedure would be realised too late. It should be emphasised that this statement does not take into account current technology (*e.g.* modern AI supported by GPUs), whose internal behaviour is largely unknown and ultimately hallucinates [10]. It can be said that humanity is in a worse position. One point to consider is the consequences of the efficient use of an unintentionally mechanised process or a hallucination (either for legitimate or illegitimate reasons).

Finally, to situate this paper discussion in a broader philosophy realm, Kant *cf.* [30] suggested that it is possible to establish *moral maxims* for guiding ethical principles. In short, a moral maxim is an idea that can be universally applied by a rational agent and whose results are always *good*. In this sense, following Peirce, *cf.* [29], moral goodness is about the aesthetics of goodness. Which, following Baumbaum, Tanizaki discusses how aesthetically ugly efficiency is, [31].

As a closure, the following quote, although widely known, is still relevant and necessary in the present times:

> *The premier demand upon all education is that Auschwitz not happen again.* [1]

6 CONCLUSION

This paper adds a possible new dimension to the dialogue about computing moral judgement. In short, in addition to the dimension of whether morality can be computed; and the dimension of how to design moral/ethical guidelines for machines to follow; this paper raises a higher order question of whether this should be done.

As discussed throughout the paper, extensive data collected on the Holocaust suggests that people, at least modern people *cf.* [5], have a tendency to delegate moral judgement to bureaucracy (perhaps as a measure of self-preservation). An analogous phenomenon has been observed in people who use AI as a co-pilot to work [12].

In addition, machines cannot be considered as responsible agents [13]. Neither are they capable of realising whether their computations are legitimate or not, as the discussion about the sorting algorithm illustrated [6]. Machines can be considered the pinnacle of bureaucracy, as all rules (no matter what rules they are) are blindly followed and with astonishing speed (for better or worse).

But if it is known that a module as such could prevent a catastrophe, even if only eventually, should a module as such be prevented from existing? This is the dilemma posed by this paper.

For future work, in addition to further assessing the suitability of the proposed dilemma based on in-depth reflection or a broader philosophical literature base, it is necessary to discuss to what extent, under what circumstances and how ethical rules can be included as dimensions of security requirements. Both from technical and policy perspectives.

ACKNOWLEDGMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project 2022.06822.PTDC (https://doi.org/10.54499/2022.06822.PTDC). The work of Pedro Oliveira was supported by the doctoral Grant PRT/BD/154311/2022 financed by the Portuguese Foundation for Science and Technology (FCT), and with funds from European Union, under MIT Portugal Program.

REFERENCES

[1] Theodor W Adorno. 1998. Education after Auschwitz. *Critical models: Interventions and catchwords* 191 (1998), 204.

[2] Isaac Asimov. 2004. *I, robot.* Vol. 1. Spectra.

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[4] Z. Bauman. 1995. *Life in Fragments: Essays in Postmodern Morality.* Wiley.

[5] Z. Bauman. 2013. *Modernity and the Holocaust.* Polity Press.

[6] E. Black. 2002. *IBM and the Holocaust: The Strategic Alliance Between Nazi Germany and America's Most Powerful Corporation.* Three Rivers Press.

[7] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576. https://doi.org/10.1126/science.aaf2654 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaf2654

[8] Maria Cramer. 2021. A.I. Drone May Have Acted on Its Own in Attacking Fighters, U.N. Says. *The New York Times* (2021). https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html.

[9] Adelino de Castro Oliveira Simões Gala. 2016. *Confrontações entre máquinas físicas, máquinas semióticas e máquinas ontológicas.* Ph.D. Dissertation. Pontifícia Universidade Católica de São Paulo.

[10] Yashar Deldjoo. 2023. Fairness of chatgpt and the role of explainable-guided prompts. *arXiv preprint arXiv:2307.11761* (2023).

[11] G. Deleuze. 1988. *Bergsonism.* Zone Books.

[12] Fabrizio Dell'Acqua. 2022. *Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters.* Technical Report. Working paper.

[13] V. Dignum. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Springer International Publishing.

[14] Edsger W Dijkstra. 1996. The next fifty years. *EW Dijkstra Archive, University of Texas, Document EWD1243a.[web page] http://www. cs. utexas. edu/users/EWD/transcriptions/E WD12xx/EWD1243a. html* (1996).

[15] Veljko Dubljević. 2020. Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics* 26 (2020), 2461–2472.

[16] Hubert Etienne. 2021. The dark side of the 'Moral Machine'and the fallacy of computational ethical decision-making for autonomous vehicles. *Law, Innovation and Technology* 13, 1 (2021), 85–107.

[17] Paul Feyerabend. 1987. *Farewell to reason.* Verso.

[18] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).

[19] Francisco S. Marcondes, Ítalo S. Vega, and Eduardo Camargo. 2019. *TECHNO.EXE.* Vol. 1. Editora Pontocom, Chapter Cognição e Computação.

[20] Masahiro Morioka. 2017. The Trolley Problem and the Dropping of Atomic Bombs. *Journal of Philosophy of Life* 7, 2 (2017), 316–337.

[21] Christine Moser, Frank den Hond, and Dirk Lindebaum. 2022. Morality in the age of artificially intelligent algorithms. *Academy of Management Learning & Education* 21, 1 (2022), 139–155.

[22] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 72–78.

[23] Merel Noorman. 2023. Computing and Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[24] Winfried Noth. 2002. Semiotic machines. *Cybernetics & Human Knowing* 9, 1 (2002), 5–21.

[25] R.S. Pressman and D. Bruce R. Maxim. 2014. *Software Engineering: A Practitioner's Approach.* McGraw-Hill Education. https://books.google.pt/books?id=i8NmnAEACAAJ

[26] Duncan Purves, Ryan Jenkins, and Bradley J Strawser. 2015. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice* 18 (2015), 851–872.

[27] Silviya Serafimova. 2020. Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–10.

[28] Michael Sipser. 2012. *Introduction to the Theory of Computation.* Cengage Learning.

[29] Catherine M Smith. 1972. The Aesthetics of Charles S. Peirce. *The Journal of Aesthetics and Art Criticism* 31, 1 (1972), 21–29.

[30] Elliott Sober. 2020. *Core questions in philosophy.* Routledge.

[31] J. Tanizaki. 2019. *In Praise of Shadows.* Random House.

[32] Alfred Tarski. 1944. The Semantic Conception of Truth: and the Foundations of Semantics. *Philosophy and Phenomenological Research* 4 (3 1944), 341. Issue 3. https://doi.org/10.2307/2102968

[33] U.S.A. Secretary of Defense 2023. *DOD DIRECTIVE 3000.09 - AUTONOMY IN WEAPON SYSTEMS.* U.S.A. Secretary of Defense.

[34] Wendell Wallach, Colin Allen, and Iva Smit. 2020. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In *Machine Ethics and Robot Ethics.* Routledge, 249–266.

[35] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *SIGCAS Comput. Soc.* 47, 3 (Sept. 2017), 54–64. https://doi.org/10.1145/3144592.3144598