MORAL IMITATION:

CAN AN ALGORITHM REALLY BE ETHICAL?

Anuj Puri*

Introduction of algorithms in the realm of public administration bears the risk of reducing moral dilemmas to epistemic probabilities. This paper explores the interlinkages between attribution of moral agency on algorithms and algorithmic injustice. While challenging some of the fundamental assumptions underlying ethical machines, I argue that the moral algorithm claim is inherently flawed and has particularly severe consequences when applied to algorithms making fateful decisions regarding an individual's life. I contend that free will, consciousness and moral intentionality are *sine qua non* for any moral agent. A well-known objection to the Turing Test is cited for the proposition that, while an algorithm may imitate morality, an algorithm cannot be ethical unless it understands the moral choices it is making. I raise a methodological objection regarding transposing moral intuitions on algorithms through global surveys. I cite the 'consciousness thesis' for the principle that without consciousness there cannot be moral responsibility. Moral justifications form the bedrock of legal defenses. In the absence of moral agency and the algorithm's inability to be held morally responsible, any attempt by the firms developing and/or deploying the algorithm to escape accountability is untenable. I highlight the grave cost of masking algorithmic injustices with ethical justifications and argue for strict liability for any firm deploying algorithms in the public policy realm.

A. Introduction

Algorithms are increasingly playing a greater role in the realm of public administration.[1] From justice dispensation[2] to predictive policing,[3] algorithms are increasingly making life-altering decisions that hitherto lay solely in human domain.[4] Winfield *et al* state that the economic and societal implications of the so-called fourth industrial revolution are ripe for political and public debate.[5] The growing influence of algorithms in the public sphere has led philosophers to explore their ethical aspects and the programmers to code ethical constraints into them. But what does it mean to be moral or ethical? Does this question have the same meaning when we ask it of humans and algorithms? Put differently, can we apply the same moral standards to algorithms that we apply to humans? Importantly, should we? This paper, while appreciative of the technological progress made in the field of Artificial Intelligence ("AI"), strikes a cautionary note against declaring algorithms ethical.

---

* PhD student at the St Andrews and Stirling Graduate Programme in Philosophy (SASP), University of St Andrews, LL.M. Columbia University School of Law. Prior to commencing my doctoral research, I practiced law before the Supreme Court of India and was a visiting faculty at NALSAR Hyderabad and NLU Nagpur. For helpful discussion and comments, I am thankful to Kirstie Ball and Rowan Cruft. I am grateful to the team at Rutgers Law Record for their editorial guidance and support.

[1] *See* Karen Hao, *AI is Sending People to Jail—and Getting It Wrong*, MIT TECH. REV. (January 21, 2019), https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/ (last visited Oct. 14, 2020); Hannah Couchman, *Policing by Machine–Predictive Policing and the Threat to Our Rights*, LIBERTY (February 1, 2019), https://www.libertyhumanrights.org.uk/issue/policing-by-machine/ (last visited Oct. 14, 2020).

[2] Hao, *supra* note 1

[3] Couchman, *supra* note 1.

[4] Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3(2) BIG DATA & SOC'Y 1, 1 (2016).

[5] Alan Winfield et al., *Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems*, 107(3) PROC. OF THE IEEE 509, 509 (2019).

To illustrate the enormous costs of introducing the linguistic trojan horse of morality[6] in the world of algorithmic computation, this paper specifically focuses on the algorithms deployed in the realm of public administration.

My analysis is driven by the concern that introducing algorithms into the realm of public administration, even those that enjoy a degree of autonomy on account of their deep learning capabilities, could reduce moral dilemmas to epistemic probabilities.[7] I challenge some of the fundamental assumptions underlying the rise of Artificial Moral Agents ("AMA"). I argue that free will, consciousness, and moral intentionality are *sine qua non* for *any* moral agent.[8] Absent the aforementioned, algorithms cannot be considered AMA. Additionally, this paper posits that we have been asking the wrong moral question of algorithms. Pertinently, when we ask whether an algorithm can be moral or ethical, we are not interested in an algorithm as a moral agent, but rather, we are seeking moral responsibility. In the absence of consciousness, algorithms cannot be held morally responsible. The absence of moral intentionality further undermines the moral claim of algorithms. I raise a methodological objection about transposing moral intuitions on algorithms through global surveys. I rely on a well-known objection to the Turing Test known to show that while an algorithm may imitate morality, unless it understands the moral choices it is making, an algorithm cannot be ethical.[9]

In conclusion, I state that providing ethical justifications for algorithmic errors has dire consequences. On account of the dangerous nature of algorithmic intervention in policy domain, the absence of moral responsibility is no bar for strict legal liability. At the present juncture of moral-technical evolution, our best course of action is imposing a strict joint and several liability on the Human and the AI. My hope behind writing this paper is that the objections raised would help the proponents of ethical algorithms better appreciate the normative challenges they face.

B. What is an Algorithm?

Algorithm, like most multi-faceted concepts, is plagued by many popular and technical definitions. Since this paper seeks to challenge the ethical feasibility of algorithms, it is important to define them and specify the types of algorithms that I will be subjecting to my ethical analysis. As per Kearns & Roth, "At its most fundamental level, an algorithm is nothing more than a very precisely specified series of instructions for performing some

[6] Noel E. Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. OF THE RED CROSS 787, 793 (2012).

[7] This is most evident in the realm of predictive policing models which deploy predictive mapping and individual risk assessment programs to predict where crime will happen and even who will commit it. *See* Couchman, *supra* note 1; *see also* Christopher New, *Time and Punishment*, 52(1) ANALYSIS 35 (1992) (arguing that the problem of pre-punishment is an epistemic problem rather than a moral one).

[8] Similar objections have been raised by Kenneth Einar Himma in context of ICTs. Kenneth Einar Himma, *Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent?* 11 ETHICS & INFO. TECH. 19, 19 (2009). The key differences between our approaches lie in Himma's extensive focus on consciousness, as well as my acknowledgment of algorithm's ability to imitate morality, my response to Moor's Ethical machines, my methodological objection, and my focus on algorithms in the public administration realm. *See id.* at 24-28.

[9] *See* John Searle, *Minds, Brains, and Programs*, 3 BEHAV. & BRAIN SCI. 417, 419-20 (1980).

concrete task."[10] Mittelstadt *et al.* state, "'Algorithm' has an array of meanings across computer science, mathematics and public discourse" and "[a]ny attempt to map an 'ethics of algorithms' must address this conflation between formal definitions and popular usage of 'algorithm.'"[11] After analysing various technical and popular definitions, they limit their analysis of ethical concerns to algorithms "that make generally reliable (but subjective and not necessarily correct) decisions based upon complex rules that challenge or confound human capacities for action and comprehension."[12]

For the purposes of this paper, I will be limiting my analysis to similar algorithms. Since I am interested in the ethical status of algorithms, I will be extending my analysis to machine learning[13] and deep learning algorithms.[14] Attribution of moral agency has implications across the entire spectrum of algorithmic uses, including whether to run a smart washing machine or predict recidivism.[15] However, depending on the assigned function, algorithmic errors may result in a variety of issues, from defective products to algorithmic injustice.[16] Although my analysis is applicable to all algorithms, this paper's focus is limited to algorithms in public administration because the assignment of moral agency to such algorithms has stark implications. With such a focus, I now turn to their ethical aspect.

C. The Ethical Claim

What does it mean to be "moral" or "ethical?"[17] The answer to this question is beyond the scope of this paper. This paper focuses on whether we can apply the same moral standards to algorithms that we apply to humans. According to Fieser, "The field of ethics (or moral philosophy) involves systematizing, defending, and recommending concepts of right and wrong behavior."[18] As outlined by Anderson and Anderson, "*Machine ethics* is concerned with giving *machines* . . . a procedure for discovering a way to resolve the ethical dilemmas

[10] MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 4 (2019).
[11] Mittelstadt, *supra* note 4, at 2.
[12] *Id.* at 3. *See generally* Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 J. ON TELECOMM. & HIGH TECH. L. 203 (2015).
[13] *See* Taiwo Oladipupo Ayodele, T*ypes of Machine Learning Algorithms*, *in* NEW ADVANCES IN MACHINE LEARNING 19 (Yagang Zhang ed., 2010) ("Machine learning is about designing algorithms that allow a computer to learn. Learning does not necessarily involves consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Thus, many machine learning algorithms will barely resemble how human might approach a learning task.").
[14] *See* Li Deng & Dong Yu, *Deep Learning: Methods and Applications*, 7 FOUND. & TRENDS IN SIGNAL PROCESSING 197, 198-202 (2014) (Deep learning algorithms are "a class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification.").
[15] *See* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES 1, 1 (2018) (highlighting the limitations of algorithmic recidivism prediction).
[16] *See* Karni A. Chagal-Feferkorn, *Am I an Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 STAN. L. & POL'Y REV. 61, 84-86 (2019).
[17] Some philosophers disagree on the meaning and use of the terms 'moral' and 'ethical'. In much of the literature on machine ethics, including this paper, they are used interchangeably. *See, e.g.*, Stephen Cave et al., *Motivations and Risks of Machine Ethics*, 107 PROC. OF THE IEEE 562, 563 (2019).
[18] James Fieser, *Ethics*, INTERNET ENCYC. OF PHIL., https://iep.utm.edu/ethics/.

they might encounter."[19] But why is the field of machine ethics emerging now? We have been working with machines for thousands of years, but until recently, we have never posed a serious query regarding their moral nature. Whether it was the hand axe, the wheel, pulleys, steam turbine or airplanes, we were concerned with safety and efficiency but never the morality.[20]

What has changed? Kearns & Roth offer two reasons. First, as it pertains to algorithms, particularly models "derived directly from data via machine learning," there is "a significant amount of agency to make decisions without human intervention[.]"; second, algorithms are "so complex and opaque that even designers cannot anticipate how they will behave in many situations."[21]

A third reason of functional morality may be added to the list. Proponents of AMA state that autonomous systems are increasingly in charge of a variety of decisions that have ethical ramifications.[22] However, this accounts only for the motivation of the moral query and not a philosophical justification. Can the moral question be answered only on grounds of inevitability? If our best answer is that an algorithm is moral because algorithm *should* be moral, we have failed the test of reasoning.

The larger question deserves a closer examination, specifically, are algorithms AMA?[23] Many philosophers have hailed the rise of AMA as imminent and inevitable.[24] Following Moor's formulation of ethical machines, all algorithms pass the test of moral agency with variable degrees on account of impact[25] or as implicit[26], explicit[27] or full ethical agents.[28] However, under Moor's formulation, only a full ethical agent possesses consciousness, intentionality, and free will.[29] I disagree with this assessment. Consciousness, intentionality

[19] MICHAEL ANDERSON & SUSAN L. ANDERSON, MACHINE ETHICS 1 (Michael Anderson & Susan L. Anderson, eds., 2011).
[20] *See generally* Joaquin Ocampo et al., *The Competitive Value of Machine Safety*, MACHINE DESIGN (Aug. 10, 2018), https://www.machinedesign.com/automation-iiot/article/21837019/the-competitive-value-of-machine-safety.
[21] KEARNS & ROTH, *supra* note 10, at 7.
[22] *See* WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 8 (2009).
[23] *See* Josè Antonio Cervantes et al, *Artificial Moral Agents: A Survey of the Current Status*, 26 SCI. & ENG'G ETHICS 501, 505 (2020) ("[A]n AMA is a virtual agent (software) or physical agent (robot) capable of engaging in moral behaviour or at least of avoiding immoral behaviour. This moral behaviour may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily.").
[24] *See* Adam Poulsen et al., *Responses to a Critique of Artificial Moral Agents*, ARXIV (2019), https://arxiv.org/pdf/1903.07021.pdf; *see also* Ryan Tonkens, *A Challenge for Machine Ethics*, 19 MINDS AND MACHINES 421 (2009).
[25] James H. Moor, *The Nature, Importance, and Difficulty of Machine Ethics*, 21 INTELLIGENT SYS., IEEE 18, 19 (2006) (noting that a machine whose functions have ethical impacts maybe called ethical-impact agent).
[26] *Id.* An implicit ethical machine's internal functions implicitly promote ethical behaviour—or at least avoid unethical behaviour.
[27] *Id.* at 19-20. An explicit ethical machine represents ethics explicitly and then operate effectively on the basis of this knowledge.
[28] *Id.* at 20. ("A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them."). *See also* Cave, *supra* note 17; Luciano Floridi & J.W. Sanders, *On the Morality of Artificial Agents*, 14 MINDS & MACHINES 349, 349 (2004) (articulating a concept of moral agent not necessarily exhibiting free will, mental states, or responsibility).
[29] Moor, *supra* note 25, at 19-20.

and freewill are *sine qua non* for *any* moral agent. For reasons stated below, I assert that while an algorithm may be capable of moral imitation, it cannot really be ethical.[30]

D. The Objections

In this section, I state my normative and methodological objections to the moral machine. This section will first provide a brief overview of the debate concerning AMA and machine ethics. I will then assert that while algorithms enjoy computational agency, they lack moral agency. I tie up the exercise of moral agency to free will, which is again lacking in algorithms. In the absence of moral consciousness, algorithms cannot be held morally responsible. The absence of moral intentionality further undermines the moral claim of algorithms. Finally, I raise a methodological objection regarding transposing moral intuitions on algorithms through global surveys.

i. Brief Overview

The idea of moral machines has attracted a great degree of support,[31] with skeptics few and far between,[32] even with its early proponents acknowledging and admitting its limitations.[33] A significant amount of the scholarship supporting the idea of AMA devotes itself to fine-tuning the process without subjecting the idea of AMA, itself, to critical assessment.[34] This is in spite of an ominous warning articulated by Sharkey in the context of autonomous robots, which is equally applicable to moral algorithms

> Anthropomorphic terms like 'ethical' and 'humane', when applied to machines, lead us to making more and more false attribution about robots further down the line. They act as linguistic Trojan horses that smuggle in a rich interconnected web of human concepts that are not part of a computer system or how it operates. Once the reader has accepted a seemingly innocent Trojan term, such as using 'humane' to describe a robot, it opens the gates to other meanings associated with the natural

[30] A similar issue has been briefly explored in the case of autonomous robots. *See* Aaron M. Johnson & Sidney Axinn, *The Morality of Autonomous Robots*, 12 J. OF MIL. ETHICS 129 (2013) (explaining briefly how an autonomous robot can mimic morality but cannot be moral); *see also* Shannon Vallor & George A. Bekey, *Artificial Intelligence and the Ethics of Self-learning Robots*, *in* ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE 338 (Patrick Lin et al. eds., 2017) (explaining ethical risks posed by AI robots).

[31] *See* Poulsen et al., *supra* note 24. *See also* Allen et al., *Why Machine Ethics?*, 21 INTELLIGENT SYS., IEEE 12 (2006); Bongani Andy Mabaso, *Computationally Rational Agents Can be Moral Agents*, ETHICS & INFOR. TECH. (2020); John Danaher, *Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism*, 26 SCI. & ENG'G ETHICS 2023 (2019); David J. Gunkel, *A Vindication of the Rights of Machines*, 27 PHIL. & TECH. 113 (2013).

[32] *See, e.g.*, Patrick Chisan Hew, *Artificial Moral Agents Are Infeasible with Foreseeable Technologies*, 16 ETHICS & INFO. TECH. 197 (2014); Aimee van Wynsberghe & Scott Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, 25 SCI. & ENG'G ETHICS 719 (2019).

[33] COLIN ALLEN & WENDELL WALLACH, ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 55-68 (Patrick Lin et al. eds., 2011) (answering some of the criticism of their work on moral machine and stating that they intended their work to be start of the discussion and not the definitive word).

[34] *See generally* Anne Gerdes & Peter Øhrstrøm, *Issues in Robot Ethics Seen Through the Lens of a Moral Turing Test*, 13 J. OF INFO., COMM. & ETHICS IN SOC'Y 98 (2015); John-Stewart Gordon, *Building Moral Robots: Ethical Pitfalls and Challenges*, 26 SCI. & ENG'G ETHICS 141 (2020); Bertram F. Malle, *Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots*, 18 ETHICS & INFO. TECH. 243 (2016); Cervantes et al., *supra* note 23.

> language use of the term that may have little or no intrinsic validity to what the computer program actually does.[35]

In the face of Sharkey's well-articulated objection, one could have assumed that the debate surrounding AMA would have proceeded with greater caution. However, since the trend continues unabated, in this paper, I side with the skeptics, and challenge some of the foundational assumptions of AMA, beginning with their claim to moral agency.

ii. Moral Agency

While algorithms may have computational agency, they lack moral agency. By virtue of their programming, algorithms possess computational agency, which means that they are fit for their designed purpose. However, a computationally correct answer is not a moral outcome. Unlike a mathematical judgment, it is the motivating force of a normative judgment which makes it normative.[36] A moral agent is defined as "a person who has the ability to discern right from wrong and to be held accountable for his or her own actions."[37] According to Wallach and Allen, "A central feature of the human experience as moral agents is that people frequently feel poised between acting selfishly and acting altruistically."[38] Bandura explains that "Moral agency has dual aspects manifested in both the power to refrain from behaving inhumanely and the proactive power to behave humanely."[39] Further, as it pertains to moral agency, algorithms do not satisfy the requirements of any of the major ethical theories such as: Utilitarianism, Kantian ethics, and Virtue ethics.[40]

An effective way of testing an algorithm's moral agency may be a variation of the Turing Test.[41] Several philosophers have debated the possibility of an Ethical Turing Test.[42] I disagree with the Ethical Turing Test's feasibility on account of the objection to the Turing Test known as the 'Chinese Room Argument'.[43] The 'Turing Test' provides that a computer should be considered intelligent if it can pass for a human in an online chat.[44] The objection

---

[35] Sharkey, *supra* note 6, at 793.

[36] Connie S. Rosati, *Moral Motivation*, STAN. ENCYC. OF PHIL. (2016), https://plato.stanford.edu/entries/moral-motivation/. A case in point is Kant's *prudent shopkeeper*, who is not motivated by a sense of duty, but who treats his customers honestly "purely on the basis of prudential calculation," so as to maximize his long-term profit. Since the shopkeeper's behaviour is "based entirely on a morally irrelevant prudential calculation, it has no genuine moral worth." Henry E. Allison, KANT'S GROUNDWORK FOR THE METAPHYSICS OF MORALS: A COMMENTARY 88 (Oxford Univ. Press, 1st ed., 2011).

[37] *Moral Agent*, ETHICS UNWRAPPED (2019), https://ethicsunwrapped.utexas.edu/glossary/moral-agent (last visited on May 7, 2020).

[38] WALLACH & ALLEN, *supra* note 22, at 61-62.

[39] Albert Bandura, *Selective Moral Disengagement in the Exercise of Moral Agency*, 31 J. OF MORAL EDUC. 101, 101 (2002).

[40] For a discussion of AMA and various moral theories, see generally Colin Allen et al., *Prolegomena to Any Future Artificial Moral Agent*, 12 J. OF EXPERIMENTAL & THEORETICAL A.I. 251 (2000); Anthony F. Beavers, *Between Angels and Animals: The Question of Robot Ethics, or Is Kantian Moral Agency Desirable?*, ASS'N FOR PRAC. & PRO. ETHICS (2009); William A. Bauer, *Virtuous vs. Utilitarian Artificial Moral Agents*, 35 A.I. & SOC'Y 263 (2020).

[41] Alan Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (1950).

[42] Winfield et al., *supra* note 5, at 513.

[43] Searle, *supra* note 9.

[44] David Cole, *The Chinese Room Argument*, STAN. ENCYC. OF PHIL. (2020), https://plato.stanford.edu/archives/spr2020/entries/chinese-room/; *see also* Turing, *supra* note 40.

proposes an exception to the Turing Test in form of a thought experiment involving a native English speaker who does not understand the Chinese language, sitting alone in a room, responding to Chinese symbols slipped under the door on the basis of instructions written in English.[45] The person in the room by following instructions in English maybe able to respond convincingly in Chinese Language without understanding the language.[46] Cole summarizes the Turing Test's inadequacy on account of this objection as, "programming a digital computer may make it appear to understand language but does not produce real understanding."[47] Similarly, I assert that an algorithm may be capable of moral imitation, but not moral agency. When an algorithm chooses five lives over one life, it does not understand the choice; it is merely following a set of instructions.

At this stage, the co-relation between free will and moral agency may be briefly addressed. It is the possibility of acting to the contrary that makes human actions moral or ethical. In the case of algorithms, this possibility does not exist. Whether we consider machine learning or deep learning, the computation would always point to one 'correct' answer. This methodology is computational, not ethical. The resultant action is a calculated one, not a moral one.

Additionally, an inquiry into the ethics or morality of algorithms is not an inquiry of moral agency, but of moral responsibility. Moral agency and responsibility are related, but fundamentally different concepts. Moral agency is the capacity of moral action, while moral responsibility is the ability to face the consequences of the action. Algorithms fail to meet the test of moral responsibility, as outlined below.

iii. Consciousness and Moral Responsibility

I disagree with Moor's hypothesis that consciousness and intentionality are necessary conditions of only fully ethical agents.[48] Consciousness and intentionality are necessary conditions for any moral agent, regardless of the degree of agency. Consciousness is an epistemic condition of moral responsibility.[49] Levy explains consciousness thesis as "the thesis that consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility."[50] He further states, "only when we are conscious of the facts that give our actions their moral significance are those actions expressive of our identities as practical agents and do we possess the kind of control that is plausibly required for moral responsibility."[51] At the present stage of technological development, an algorithm is not conscious of the 'moral choices' it is making. In the

[45] Cole, *supra* note 44.
[46] Larry Hauser, *Chinese Room Argument*, INTERNET ENCYC. OF PHIL., https://iep.utm.edu/chineser/ (last visited Nov. 2, 2020).
[47] Cole, *supra* note 44.
[48] *See* Moor, *supra* note 25.
[49] *See* Fernando Rudy-Hiller, *The Epistemic Condition for Moral Responsibility*, STAN. ENCYC. OF PHIL. (2018), https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic.
[50] NEIL LEVY, CONSCIOUSNESS AND MORAL RESPONSIBILITY 1 (2014).
[51] *Id.*

absence of consciousness, algorithms cannot be held morally responsible.[52] However, as I subsequently argue, the absence of moral responsibility is not a bar to legal liability. But before that, I will briefly state my final normative objection to Moor's ethical formulation.

iv. Moral Intentionality

In *Intentionality*, Pierre provides that "[I]ntentionality is the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs."[53] Killen and Rizzo explain that "Moral judgments require the recognition of intentionality, that is, an attribution of the target's intentions towards another."[54] Further, Himma states "Agency, as a conceptual matter, is simply the capacity to cause actions – and this requires the capacity to instantiate certain intentional mental states."[55] An algorithm lacking moral intentionality thus is incapable of making moral judgments and lacks moral agency. In the absence of intentionality and consciousness, an algorithm cannot be moral or ethical.

v. Methodological Objection

In this section, I analyze the Moral Machine[56] project in order to raise a methodological objection against conferring moral status on machines. The Moral Machine project proposed a series of choices, framed in the form of trolley problem, which arise from the use of autonomous vehicles.[57] The global ethical study, which was the first of its kind, drew millions of responses in multiple languages from across the world.[58] The Moral Machine project drew overwhelming utilitarian response globally, specifically: "People generally opted to save more lives when possible, children rather than adults, and people rather than animals."[59] When posed with variations of the query, whether a self-driven car should kill fewer people in order to save more, an overwhelming number of the participants answered in the affirmative.[60] However, the participants shifted to a more deontological outlook when it came to their and loved ones safety as passengers.[61] The responses also varied depending

[52] *See* Rajakishore Nath & Vineet Sahu, *The Problem of Machine Ethics in Artificial Intelligence*, 35 A.I. & SOC'Y 103 (2020) ("The notion of mind is central to our ethical thinking, and this is because the human mind is self-conscious, and this is a property that machines lack, as yet.").

[53] Jacob Pierre, *Intentionality*, STAN. ENCYC. OF PHIL. (2019), https://plato.stanford.edu/archives/win2019/entries/intentionality.

[54] Melanie Killen & Michael T. Rizzo, *Morality, Intentionality, and Intergroup Attitudes*, 151 BEHAV. 337, 337 (2014).

[55] Himma, *supra* note 8, at 20-21.

[56] The Moral Machine project at Massachusetts Institute of Technology ("MIT") served as "a platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars." *See* Iyad Rahwan, *Moral Machine*, MIT MEDIA LAB (last accessed Oct. 28, 2020), https://www.media.mit.edu/projects/moral-machine/overview/.

[57] Darius-Aurel Frank *et al*, *Human decision-making biases in the moral dilemmas of autonomous vehicles*, 9: 13080 SCI REP 1 (2019).

[58] *See* Chris O'Brien, *Moral Machine is Learning How We Want Self-Driving Cars to Kill*, VENTUREBEAT (October 15, 2019), https://venturebeat.com/2019/10/15/moral-machine-is-learning-how-we-want-self-driving-cars-to-kill/.

[59] *Id*.

[60] *See* Edmond Awad et al., *The Moral Machine Experiment: 40 Million Decisions and the Path to Universal Machine Ethics*, 563 NATURE 59 (2018).

[61] Frank *et al*, *supra* note 57, at 1.

upon the age of the person in danger.[62] These varied moral intuitions were used to train the machine learning algorithms that would operate the autonomous vehicles.[63]

Autonomous vehicles are a stark example of reduction of moral dilemmas to epistemic probabilities. While discussing the dilemmas of autonomous ethics in cases where the autonomous vehicle has to decide the course of action in a situation involving decreasing the risk to its passengers by increasing the risk to a potentially larger number of non-passengers, Shariff *et al* state, "[W]hile these decisions will most often involve probabilistic trade-offs in small-risk manoeuvres, at its extreme the decision could involve an autonomous vehicle determining whether to harm its passenger to spare the lives of two or more pedestrians, or vice versa."[64]

Significantly, the Moral Machine project does not address the fundamental query of how we can transpose our ethical intuitions on non-human actors. For instance, we can apply utilitarian considerations to animals because animals are sentient beings; or, we can apply a Kantian perspective to animals because we can be concerned with how our treatment of animals can affect our duties to persons.[65]

However, as it pertains to machines, we are trying to program them in an anthropomorphic, ethical way.[66] The question is whether we can apply utilitarian principles to non-sentient beings? Can we term algorithms as AMA and continue to deploy ethical standards qua them solely from human perspective? From the machine's perspective, is the correct formulation of the utilitarian question whether to save five lives or one? Or, is it whether to save the expensive car or the cheaper car? From a deontological ethics perspective, it seems paradoxical to treat an algorithm as a means to an end and then expect it to satisfy the test of Kantian ethics.[67] Lastly, from a virtue ethics perspective, how do we determine the virtuous mean? What would constitute *eudaimonia or phronesis* from an autonomous vehicle's perspective?[68] The moral machine survey makes a fatal error in equivocating human ethical intuitions with constraints that need to be programmed into autonomous vehicles. Safety features, no matter how laudable, are no substitutes for ethics. Apart from the normative objection, there is a methodological concern as well as regards to how the moral status of algorithm is being conceived.

---

[62] *Id*. at 8.
[63] *Id*.
[64] Azim Shariff *et al*, *Psychological roadblocks to the adoption of self-driving vehicles*, 1 NAT HUM BEHAV 694 (2017).
[65] *See* Lori Gruen, *The Moral Status of Animals*, STAN. ENCYC. OF PHIL. (2017), https://plato.stanford.edu/archives/fall2017/entries/moral-animal/.
[66] *See* Karni Chagal-Feferkorn, The Reasonable Algorithm, UNIV. ILL. J. OF L., TECH. & POL'Y 111, 130 (2018) (discussing the conceptual difficulties stemming from applying a "reasonableness" standard to nonhumans, including the intuitive reluctance of subjecting non-humans to human standards).
[67] Tonkens states that the development of Kantian Artificial Moral Agents is anti-Kantian. Tonkens, *supra* note 24, at 429.
[68] Rosalind Hursthouse & Glen Pettigrove, Virtue Ethics, STAN. ENCYC. OF PHIL. (2016) https://plato.stanford.edu/entries/ethics-virtue/ (*[E]udaimonia* "is standardly translated as 'happiness' or 'flourishing' and occasionally as 'well-being.'"). James Tiles, MORAL MEASURES: AN INTRODUCTION TO ETHICS WEST AND EAST 97 (Routledge, 2000) ("*[P]hronesis* involves the ability to deliberate well, i.e. settle on a course of action in the light of a goal, not merely in the light of some particular goal that may be before one, but in the light of the goals that contribute to a good life in general.")

As early as 1950, Turing cautioned against a statistical survey approach to answer questions relating to machine thinking:

> I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd.[69]

The normative objection is germane to our analysis, even for machine ethics. We cannot collectively will our way to moral machines. Much like the question, 'Can machines think?', the question 'Are machines moral?' cannot be answered through a statistical survey. Transposing globally collected ethical intuitions as data for machine learning algorithm may give rise to moral imitation but in the absence of the pre-requisites highlighted in this paper, an algorithm cannot be said to be ethical.

E. Conclusion

The idea of an ethical algorithm has an intuitive appeal. People are likely to feel better knowing that momentous decisions concerning their lives are being made only after due ethical considerations. Perhaps, it is our distinct unease of losing control over our public policy affairs that has led us down the wrong path of concluding that algorithms can be ethical. The question we should be asking ourselves, but perhaps are reluctant to ask: "Is there is any ethical justification when someone is denied social benefits[70] or sentenced incorrectly or not hired or denied loans[71] or racially discriminated in search engines[72] on account of an algorithmic error?" The aforesaid illustrates the real cost of transplanting an ethical mask on artificial intelligence. When we cloak cold algorithmic calculations under the warm ethical cover, we mask costly errors that have human repercussions. Constraints programmed into a code are safety features and not ethics. An AI system designed to act as a lifeguard would not have to sacrifice "anything of comparable moral importance" to save a drowning child.[73] If it succeeds, it would be considered fit for purpose and not heroic. And if it fails, the failure would be considered a mechanical error and not tragic.

The issue of ethical algorithms has significant legal implications. Moral and ethical justifications form the basis of legal defences. An AMA can be successfully used by a corporation to mitigate or even abdicate legal liability.[74] It can be argued that a deep

[69] Turing, *supra* note 41.
[70] *See* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 59 (2018).
[71] *See* Joy Buolamwini, *Algorithmic Justice League*, MIT MEDIA LAB (2019), https://www.media.mit.edu/projects/algorithmic-justice-league/overview/ (last visited on April 26, 2020).
[72] *See* Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, 7 (May 2014) https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
[73] *See* Peter Singer, *Famine, Affluence, and Morality*, 1 PHIL. & PUB. AFFS. 229, 231-33 (1972).
[74] *See* Cathy O'Neil, *Amazon's Gender-Biased Algorithm Is Not Alone*, BLOOMBERG (Oc. 6, 2018, 9:00 AM), https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone.

learning algorithm as an AMA was sufficiently autonomous to be legally held responsible for the error. When algorithms enter the public administration realm, we are no longer dealing with a defective product but algorithmic injustice. The inapplicability of moral agency to an algorithm takes away much of the moral justification that may be offered by a corporation in form of plausible deniability. Further, the absence of moral responsibility is not a bar for legal liability. The age old principle of strict liability when dealing with an inherently dangerous object is squarely applicable to Artificial Intelligence when it enters policy domain.[75] Depending on the type of algorithm, whether it is deep learning or controlled by a programmer, the liability maybe joint or several.[76] Artificial consciousness, when achieved, maybe moral but AI is certainly not. I am not ruling out the possibility of Artificial Moral Agents in perpetuity, but the current state of affairs necessitates an ethical code for programmers and legal liability for corporations deploying algorithms in the policy realm.

[75] *See* Rylands v. Fletcher [1868] UKHL 1; *see also* Chagal-Feferkorn, *supra* note 16, at 79-80.

[76] For an exposition of shared moral responsibility with robots, see Gordana Dodig-Crnkovic & Daniel Persson, *Sharing Moral Responsibility with Robots: A Pragmatic Approach*, 173 FRONTIERS IN A.I. & APPLICATIONS 165 (2008).