Epistemology and *Algo-reliabilism*: A Pathway to Sound Ethical Artificial Intelligence

Helen Titilola OLOJEDE, PhD, (SCI) [2] Candidate

Department of Philosophy, National Open University of Nigeria.

helenolojede@gmail.com

holojede@noun.edu.ng

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: Ethics, AI, Algoreliabilism, Goods, Epistemology

Extended Abstract

The birth of AI as a field is usually linked to the 1956 conference with figures involving Marvin Minsky and John McCarthy. At the wake of AI was the debate between the proponents of weak and strong AI regarding whether the technology would displace/replace humans. One will not be incorrect to say at this juncture in the debate that more concerted efforts on the parts of various stakeholders is to ensure the safety of the use of AI for humans and that it better serves humanity. To this end, artificial intelligence is a topical issue in the world and in academic discourse. Its influence spans various fields of endeavors such as healthcare, education, workplace etc. No doubt, it holds great prospects even as it is revolutionizing our world. Quite important is also the need for ethical regulation which many scholars, academic endeavors, regional and international organizations, ethical AI councils are already responding to. On the one hand, some of the ethical challenges posed by AI are not limited to: autonomy versus altering the society in ways that might not be agreeable to others, safety and uncertainty, privacy. There are also issues that revolve around bias or discrimination which include for instance, facial recognition, medical records, weaponized drones, historic injustice etc. On the other hand, some of the ethical principles already proposed to guide AI include: transparency, inclusion, responsibility, impartiality, reliability, security. However, a lot more still needs to be done. There is no doubt from the above that there are lots of on-going efforts to ethically regulate AI. Nonetheless, there

is a need to clarify certain questions regarding AI ethics. To whom or to what does AI ethics apply? Is ethics or ethical principles for the originators of AI or for AI itself or for the users of AI? How should we understand individual ethics questions? How about collective or societal ethics question? This paper thus, tries to clarify some of these questions as it argues that AI ethics cannot be devoid of an understanding of its epistemology and in turn the agency of the efficient cause. In other words, sound ethical AI ought to incorporate the epistemology of algorithms. Thus, while logic and probability, which are branches of Mathematics and aspects of Philosophy underlie and play significant roles in the development of AI; algorithm relies on and employs some input and with the help of Mathematics and Logic bring about the output. AI algorithm however, makes use of both the inputs and outputs concurrently for it to 'learn' the data and generate output when new inputs are given. Based on the importance algorithms play in the life of AI system, this paper argues for what it calls *algorithmic reliabilism* in the development of AI.

The reliabilism in 'Algorithmic reliabilism' is an offshoot of the traditional reliabilist epistemologies, which states that, one knows a particular proposition, if one believes the proposition and if the proposition is true. It however has closer affinity with the more recent process reliabilism which emphasizes the importance of carrying out a set of actions or procedures in order to arrive at an intended or desired outcome. Consequently, 'Algorithmic reliabilism' is conceived as a set of normative accountability mechanisms, one that incorporates certain Goods in order to make AI safer and more just while becoming more efficient for persons as individuals and groups of persons. By Good, we mean for instance, that an apparatus or equipment is good if it serves us in the aim for which it was envisioned. This means, it is good as a result of its efficiency in achieving a desired outcome. The outcome may either be desired for itself, or it could be that we seek it as a means to some underlying end. If we seek it for itself, we tend to think of it as a good and so desire it, in and of itself. If, however, we seek it for an underlying end, that end becomes the good. The sequence of means and ends either go on indefinitely or it has to stop when we arrived at some anticipated object(s) which are ends in themselves.

More concretely, I conceive Good in the context of AI ethics as the output of *algorithmic reliabilism*. These Goods should include: Solidarity, human dignity, interpretive fineness and

natural law.

Solidarity – summarily, in AI ethics context, Good means that we should look out for one another both as individuals and as groups in terms of inclusion, diversity, transparency. Every member of the larger society be it AI designers, regulators, regional governments, non-profit organization, various stakeholders is expected to work together for, and commit to, the ideals that unify, uphold and sustain the public good of the larger society. It exemplifies the need to recognize the complex relationship that exists between humans and social groups.

Human dignity – AI systems should be such that it does not ride rough-shod on humans. I know this is quite broad and needs some unpacking however, a sound ethical AI cannot but be safe and respectful of the human agency. Human dignity is thus, the idea that a being has an intrinsic right to be valued and respected. It also means that each and every individual person is a whole, an entity in itself; and, at the same time, that everybody is part and parcel of something that goes beyond ourselves; something greater and larger than ourselves and AI developers, its deployment and ethics ought to take cognizance of the dignity of humans.

Interpretive fineness - The maxim, 'other things being equal, simpler theories are better' is one of the ideas that explain what I mean by interpretive fineness of AI system. By interpretive fineness of AI, I also mean that the algorithm that birth the AI is a complex mix of a number of ideas (not in order of importance) such as the simplicity of algorithm, its use and how easily explainable, understandable, deployable, ethically viable it is. Aristotle's assertion in *Posterior Analytics* lends credence to AI simplicity: "we may assume the superiority *ceteris paribus* of the demonstration which derives from fewer postulates or hypotheses."

Coherence – this has to do with congruence, internal consistency or cohesion that exists in AI systems, or group of systems and intended audience. Coherence has to do with how well AI and human fit well, interact seamlessly especially in an ethical manner. This is not just a consideration of humans but also of robots especially humanoids. A smooth techno/human relation with limited or mitigated ethical concerns.

SELECTED REFERENCES

Sven Nyholm. 2020. Human and Robots. Ethics, Agency and Anthropomorphism. London: Rowman & Littlefield, 2022

AI Bill of Rights. Making Automated Systems Work For The American People. The White House, October 2022.

Goldman, A. 2011. A Guide to Social Epistemology. *Social epistemology: Essential Readings.* Goldman, A. & Whitcomb, D. eds. Oxford University Press.

Paolo Benanti. 2020. Algo-ethics: Artificial Intelligence and Ethical Reflection Jason Borenstein and Ayana Howard. 2020. Ethical Issues in AI and the Need for AI Ethics Education. *Springer.*

Daniel Tigard. 2020. Responsible Ai and Moral Responsibility: A Common Appreciation. *Springer*

Kelvin LaGrandeur. 2020. How safe is our reliance on AI, and should we regulate it? *Springer*