Epistemic Doom In The Deepfake Era

Nadisha-Marie Aliman[1][0000−0003−3049−9327]

Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands
nadishamarie.aliman@gmail.com

Abstract. This epistemic project examines an understudied existential risk emerging in the deepfake era: the fortunately up to this time (but not indefinitely so) reversible peril of humanity's *epistemic self-sabotage* through an overestimation of algorithms linked to quantitative aspects and a paired underestimation of the own epistemic potential whose manifestations are in principle expressible via scientifically analyzable but currently often neglected qualitative facets. This scenario is metaphorically referred to as "π-Doom scenario". Instead of carefully crafting opaque hypotheses and formulating probabilistic predictions of timelines that could easily evade scrutiny and/or could even themselves be initially generated *by* algorithms, scientifically and responsibly counteracting π-Doom requires new better *riskier* explanatory theories of intelligence, creativity and consciousness. A brief non-comprehensive literature review suggests that quite a few scientific instantiations (all of which have *not* yet been provisionally refuted) of that transformative requirement are *already* available at present. In short, π-Doom – the voluntarily or unintentionally effectuated cognitive resignation engendering an irrational algorithmically-motored process of running around in (epistemic) circles that could endanger the immediate future of a vulnerable civilization like present-day humanity – is wholly contingent on a choice of focus.

Keywords: Epistemology · Intelligence · Creativity · Consciousness .

1 The Problems

Words can be enacted as weapons of mass destruction and the linguistic definitions one uses can sometimes be decisive to the course of history. Firstly, in the midst of the current fragile ecosystem of international security including its links to cybersecurity [13], it appears straightforwardly cognizable that overhasty and clearly *misguided* algorithmic superintelligence achievement claims (motivated e.g. purely by commercial goals) could unnecessarily risk to fuel conflicts compromising world peace. Secondly, an unintentional fundamental overestimation of algorithms by the wider public could arguably, when crossing a certain threshold of exaggeration such as e.g. via narratives of qualitatively superintelligent "God-like" algorithm realizations, bring about widespread mental health issues up to a superfluous epistemic panic with economical consequences and further unpredictable second-order harms. Thirdly, against the background of these two mentioned different risk clusters, it is easily conceivable that potential malicious

humans actors deliberately attempting to utilize narrow algorithms to harm humanity (e.g. via large scale cyberattacks or robotic engines deployed in real-world settings) could be underestimated and go unnoticed while proactive defense measures requiring the vigilance of people could stay undervalued. Fourthly, in the immediate near-term, an over-reliance on algorithms encompassing the usage of algorithmic tools misconceivedly deemed to approach superintelligence (and confusingly called "AI agents") in *safety-critical* contexts that would omit to locally encapsulate those in local units controlled by people, could itself lead to existential risks for a civilization since the trade-off-based combination of extremely lower latency but only even slightly lower reliability in comparison to people could yield to an impediment of human rational evaluation with lethal side-effects. Fifthly, a fundamental overestimation of algorithms which did as a reaction often unnecessarily engender their underestimation is connected to a counterproductive two-sided complex of unintentional anthropomorphization and "animization" of algorithms on the one hand and of both unintentional dehumanization of human statistical outliers and unintentional "deanimization" of non-human animals on the other hand. The latter reprents an epistemic obstacle to a more robust scientific evaluation of coming algorithmic superintelligence achievement claims. Sixthly, many of the proponents of algorithmic superintelligence implementation immanency (which is considered to be impossible given the currently best scientific explanations as can be extracted from Section 2) seem to be largely in denial of the logical impossibility of crafting/controlling an entity that would genuinely appear to be superintelligent in relation to oneself [3] and the person-hood status of such a hypothetical entity that would be categorized as better in all aspects of interest to humans from all domains of life. Left aside the point that it is impossible for the less intelligent entity to build such an entity in the first place, to frame it as a desirable controllable *product* and not as a free person shows up the underlying absolute power fantasies.

2 Possible Theoretical Solutions

Consistent with Popper's critical rationalism [30] and its recent refinements [10, 9] and adaptations to the challenges of the deepfake era [1–3], the presently best solutions to thwart the π-Doom scenario (some of whose features have been described in Section 1) reside in the creation of new more clearly formulated riskier theories able to better explain the nature of intelligence, creativity and consciousness such that an experimental problematization could be possible. In the following, a non-comprehensive enumeration of already existing algorithm-related *impossibility statements* (since those are the riskiest) is displayed. Since none of those has been provisionally refuted yet, it is rational to refrain from spreading algorithmic superintelligence achievement claims without first obligatorily having *both*: 1) attempted an experimental problematization via building a candidate algorithmic superintelligence which taken alone would be insufficient *and* 2) provisionally refuted *all* mentioned currently best theories via providing a new even better scientific theory of intelligence, consciousness and creativity.

2.1 Exemplary Algorithm-Related Impossibility Statements

The following compact non-comprehensive enumeration displays impossibility statements that are either explicitly formulated by the authors or directly entailed by their respective work. The impossibility statements stem from a wide variety of disciplines and specific research areas including e.g. biology [15, 31, 33], physics [26, 29, 32], hardware verification [19], cybernetics [25], mathematics [20], complexity science [14, 23], philosophy of science [11, 36], epistemology [1], possibility studies [6] and "cyborgnetics" [1–3].

1. Impossibility of present-day ASI implementation [32]
2. Impossibility of omnipotent algorithmic general intelligence [25]
3. Impossibility for present-day AI to create new scientific theories [36]
4. Impossibility of algorithmic general intelligence [31]
5. Impossibility of algorithmic general intelligence [20]
6. Impossibility of genuine algorithmic intelligence [15]
7. Impossibility of algorithmic consciousness [11]
8. Impossibility of present-day AI consciousness [19]
9. Impossibility of understanding by classical AI [29]
10. Impossibility of classical simulation of human mind [26]
11. Impossibility for algorithms to predict human creative leaps into the previously "impossible" [6]
12. Impossibility of non-biological agency [14]
13. Impossibility of non-organic general intelligence [23]
14. Impossibility of genuine inorganic intelligence [33]
15. Impossibility for x to reliably build entity y that is EB-measured to be superintelligent in relation to x [2, 3]
16. Impossibility of EB-measurable algorithmic general intelligence, creativity, consciousness [2, 3] (incl. impossibility for present-day AI to understand EBs [1])

3 Outlook

3.1 Ad-Hoc Maneuvers

Some companies are engaging in epistemic ad-hoc maneuvers to update their carefully formulated opaque definitions of algorithmic general intelligence (AGI) and algorithmic superintelligence (ASI). There are claims that specific levels of AGI have already been achieved and that an ASI which would be an algorithm that surpasses humans in all *economically* valuable tasks is approaching. While proponents prophesy that the latter will be achieved in a few years, it is worth repeating that for epistemic reasons, scientific claims of algorithmic superintelligence achievement claims must take present-day human civilization as whole as baseline. It must in principle include all *scientifically* analyzable tasks of interest to humanity irrespective of whether a task is widely considered to be economically valuable at present. Firstly, what is economically valuable is

subject to change and has always been. Similarly, the tasks of interest to humanity are highly adaptable and there is no reason to assume that humanity could not one day surpass its current epistemic situation via transformative creative leaps [6] in the previously thought "impossible". The ignorance of the relativity of instantied consciousness is reflected in the efforts of certain entities to frame "human-level" intelligence as an absolute stage that can be surpassed via algorithmic strategies. Secondly, more notoriously, the idea that *all* conceivable economically valuable tasks *at the present point in time* can already be automated is a heavy misconception. For example, humanity has long been interested in potentially higher developed alien civilizations that could exist elsewhere in the universe. Essentially, in this context, *multiple* tasks of interest [12, 21, 17] which *currently* appear elusive to human civilization (even though tentative hypotheses on how to achieve those one day – which are necessarily limited by the current epistemic situation of science – are repeatedly generated) have been formulated all of which would already be considered to be economically valuable *nowadays* (see also [3]). Thirdly, humanity still scientifically struggles with self-chosen tasks of interest such as nuclear fusion, quantum computing, the design of much more energy efficient batteries and so forth. Recently, a company claimed to now have created *algorithmic* life – a scientific oxymoron based on a misconception of self-replication (see e.g. [28] for a detailed better grasp). In Germany, one sometimes remarks: *Es ist fünf vor zwölf*; in light of the above, one can now even state: "Es ist fünf *nach* π-Doom!"

3.2 Peaceful Non-Algorithmic Research

Thus, even from an economic perspective, given the currently best explanations, it is irrational to claim that algorithmic superintelligence able to surpass human civilization at all tasks of interest to humanity can be implemented by human civilization or its algorithms. Concerning the opaque recursive self-improvement hypothesis, where a narrow algorithm self-transforms into an ASI, in light of diverse scientific explanations some of whose impossibility statements have been summarized in Section 2.1, one must currently conclude that it is impossible for an algorithm to reliably create those genuinely transformative creative leaps that the new scientific theories able to elevate human civilization as a whole to higher epistemic levels would entail. This does not logically exclude the possibility for a civilization that is much more advanced than present-day humanity [3] to one day indirectly build a *non*-algorithmic general intelligence [31]. Thus, a renewed *non*-algorithmic general intelligence (abbreviated *NaGI*) dream could be rationally pursued by interested parties, but those would inherently strive to simultaneously epistemically elevate human civilization as a whole by considering the problems of interest to humanity and would thus refrain from impeding world peace. Instead of harnessing misdirection strategies known from the neuroscience and psychology of magic [4, 22] and misguidedly claiming a speedy crafting of superintelligent algorithms as closed-source products, the slower goal of sincere *NaGI* dreamers is *open*, *multiple* transformative steps away, the timeline is fundamentally *unpredictable*, but *could* lead to (re)new(ed) conscious beings.

4 Conclusion

This paper – written purely for the purpose of a self-educational epistemic *art* project serving as ephemeral mental clipboard – explained why the π-Doom scenario is contingent on the (self-)inflicted informational salience of algorithmic supremacy claims regularly circulating in the information ecosystem of present-day humanity. While the misconceived overestimation of present-day algorithms has been thoroughly analyzed in recent works (see e.g. [5, 7, 16, 27, 34]), the need for a scientific paradigm shift [6, 18] where i.a. *non*-algorithmicity is key has been recently emphasized [8, 24, 31]. Will π-Doom *unnecessarily* act as *an* epistemic existential filter for humanity (i.e. one of the many possible ones, see also e.g. [3]) or will humanity free itself from the current self-imposed algorithmic redundancy and regain epistemic agency? Without the latter, one would further fail to build the augmentative algorithmic *tools* (locally encapsulated in human-controlled units) needed for the future. The π-Doom art project encodes not only *generic* conscious invariance [3], but also the idea that no algorithmic, necessarily *relative* "existential risk" (a non-obvious term [35]) can reliably delete its existence. Will the more sober *NaGI* dream (see Section 3.2) volatilize π-Doom?

References

1. Aliman, N.M.: Cyborgnetics – The Type I vs. Type II Split. Aliman, Nadisha-Marie (2021)
2. Aliman, N.M.: Cyborgnetic Invariance. Aliman, Nadisha-Marie (2023)
3. Aliman, N.M.: Acentric intelligence. PhilPapers (2024)
4. Aliman, N.M., Kester, L.: AI-Related Misdirection Awareness in AIVR. PhilPapers (2023)
5. Altmeyer, P., Demetriou, A.M., Bartlett, A., Liem, C.: Position Paper: Against Spurious Sparks-Dovelating Inflated AI Claims. arXiv preprint arXiv:2402.03962 (2024)
6. Corazza, G.E.: Beyond the adjacent possible: On the irreducibility of human creativity to biology and physics. Possibility Studies & Society **1**(1-2), 37–45 (2023)
7. Duarte, T., Barrow, N., Bakayeva, M., Smith, P.: The ethical implications of AI hype. AI and Ethics pp. 1–3 (2024)
8. Faggin, F.: Irreducible: Consciousness, Life, Computers, and Human Nature. John Hunt Publishing Limited (2024)
9. Frederick, D.: Against the Philosophical Tide: Essays in Popperian Critical Rationalism. Critias Publishing (2020)
10. Frederick, D., et al.: Falsificationism and the Pragmatic Problem of Induction. Organon F **27**(4), 494–503 (2020)
11. Garrido-Merchán, E.C.: Machine Consciousness as Pseudoscience: The Myth of Conscious Machines. arXiv preprint arXiv:2405.07340 (2024)
12. Gray, R.H.: The extended Kardashev scale. The Astronomical Journal **159**(5), 228 (2020)
13. Humphreys, D., Koay, A., Desmond, D., Mealy, E.: AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI and Ethics pp. 1–14 (2024)

14. Jaeger, J.: Artificial intelligence is algorithmic mimicry: why artificial "agents" are not (and won't be) proper agents. Neurons, Behavior, Data analysis, and Theory pp. 1–21 (2024)
15. Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., Walsh, D.: Naturalizing Relevance Realization: Why agency and cognition are fundamentally not computational. Frontiers in psychology **15**, 1362658 (2024)
16. Kambhampati, S.: Can large language models reason and plan? Annals of the New York Academy of Sciences **1534**(1), 15–18 (2024)
17. Kardashev, N.S.: Transmission of Information by Extraterrestrial Civilizations. Soviet Astronomy, Vol. 8, p. 217 **8**, 217 (1964)
18. Kauffman, S.A., Roli, A.: A third transition in science? Interface Focus **13**(3), 20220063 (2023)
19. Kleiner, J., Ludwig, T.: If consciousness is dynamically relevant, artificial intelligence isn't conscious. arXiv preprint arXiv:2304.05077 (2023)
20. LANDGREBE, J., SMITH, B.: Intelligence. And what computers still can't do. Cosmos+ Taxis **12** (2024)
21. Loeb, A.: Interstellar: The Search for Extraterrestrial Life and Our Future in the Stars. HarperCollins Publishers (2023), https://books.google.nl/books?id=BBmPzwEACAAJ
22. Lupetti, M.L., Murray-Rust, D.: (Un) making AI Magic: A Design Taxonomy. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–21 (2024)
23. MARTINELLI, E.: Complexity and Particularity: An Argument for the Impossibility of Artificial Intelligence. Cosmos+ Taxis **12** (2024)
24. Mogi, K.: Artificial intelligence, human cognition, and conscious supremacy. Frontiers in Psychology **15**, 1364714 (2024)
25. Mueller, M.: The Myth of AGI. Internet Governance Project (2024)
26. Muraleedharan, A.: Turing Machines cannot simulate the human mind. arXiv preprint arXiv:2207.05700 (2022)
27. Narayanan, A., Kapoor, S.: AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton University Press (2024)
28. Noble, R., Noble, D.: Understanding living systems. Cambridge University Press (2023)
29. Penrose, R.: Is Conscious Awareness Consistent with Space-Time Descriptions? In: Philosophy, Mathematics and Modern Physics: A Dialogue, pp. 34–47. Springer (1994)
30. Popper, K.: Conjectures and refutations: The growth of scientific knowledge. Routledge (1962)
31. Roli, A., Jaeger, J., Kauffman, S.A.: How organisms come to know the world: Fundamental limits on artificial general intelligence. Frontiers in Ecology and Evolution **9**, 806283 (2022)
32. Stiefel, K.M., Coggan, J.S.: The energy challenges of artificial superintelligence. Frontiers in Artificial Intelligence **6**, 1240653 (2023)
33. Svensson, J.: Artificial intelligence is an oxymoron: The importance of an organic body when facing unknown situations as they unfold in the present moment. AI & society **38**(1), 363–372 (2023)
34. Vallor, S.: The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking. Oxford University Press (2024)
35. Varshney, K.R.: Decolonial AI Alignment. arXiv preprint arXiv:2309.05030 (2023)
36. Velthoven, M., Marcus, E.: Problems in AI, their roots in philosophy, and implications for science and society. arXiv preprint arXiv:2407.15671 (2024)