Science in the Age of Algorithmic Chrysopoeia?

Nadisha-Marie Aliman[1][0000−0003−3049−9327] and Leon Kester[2][0000−0002−8565−3902]

[1] Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands
[2] TNO, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands
leon.kester@tno.nl

Abstract. This commentary analyzes different past approaches to scientific evaluation frameworks given algorithmic superintelligence claims and discusses guidelines for more rigorous solutions avoiding overhasty and damaging epistemic self-sabotage tendencies. Effects of spreading the misconceived idea of the *epistemic perpetuum mobile* are elucidated.

Keywords: Epistemology · Intelligence · Creativity · Consciousness.

1 The Problem

At the beginning of the deepfake era, the "AI" debate faced a precarious epistemic situation i.a. due to the foreseeable insufficiency of the mainstream empiricist paradigm [9, 10] which may *not* be fit-for-purpose for more rigorous scientific analyses of repeated algorithmic superintelligence immanency claims given that the paradigm is prone to "epistemic anarchy" [24, 26, 36] characterized by motifs such as a "post-truth world" [38] or a "post-epistemic world" [34]. Since then, an epistemic crisis became palpable and is revealed in the saliency of misguided algorithmic supremacy immanency claims circulating in the information ecosystem against which humanity did not yet develop epistemic resiliency. (The latter is linked to the so-called π-Doom scenarion [7].) In the deepfake era, an empiricism- (instead of explanation- [28, 50]) based epistemology with the epistemic aim to obtain justified and truer beliefs via probabilistic belief updates given forgeable empirical "evidence" can be an obstacle to the evaluation of algorithmic supremacy claims since the scientist unintentionally risks to offer an easily empirically-malleable attack surface that can be exploited by malicious actors e.g. equipped with knowledge from the psychology and science of magic [12, 40, 47]. Moreover, epistemic biases ranging from anthropomorphization [48] and animization of algorithms over dehumanization and deanimization of humans and non-human animals to the nadir of what one could call self-dehumanization or also colloquially "self-zombification" call for an explanation-anchored approach to mitigate self-misguidance. On the whole, the practice of the widespread empiricist paradigm may be at risk of prematurely concluding to have corroborated algorithmic supremacy (i.e. people could be prematurely convinced that the primacy of consciousness would have been provisionally refuted) – which may be i.a. highly detrimental to science [61] in the deepfake era. The latter would foster epistemic stasis and neglect the risks of *misuse* of and *over-reliance* [33] on algorithms. While there is a reluctance against strict *open* scientific evaluation frameworks for algorithmic superintelligence claims, it is important to note that given certain salient associations such as conceiving of hypothetical immanent superintelligent algorithms as "aliens", an "alien species" [53, 54] with "God-like" [56] properties at risk of *making the entire human civilization obsolete*, it is worthwhile establishing clear standards that do not engender cry wolf effects. Conveniently, past SETI research *implicitly* already analyzed *scientifically* plausible notions of what would count as a "superintelligent" civilization in relation to current humanity. Indeed, SETI research has been linked to the concept of the *cosmic mirror* which is connected to the conjecture that "*[...] SETI might unify the world because it helps human beings to see themselves in a cosmic context"* [19]. In this vein, the requirements for more rigorous scientific evaluation frameworks of algorithmic superintelligence achievement claims in the deepfake era collated in the following Section 2, incorporate explanation-anchored epistemic levels that are directly inspired by SETI scales. This procedure of introducing SETI-inspired tasks of interest (which *are* a scientifically analyzable subset of the tasks of interest to the entire human civilization) into the scientific evaluation of algorithmic supremacy claims in the deepfake era has been introduced earlier by Aliman [4, 6, 8]. Recently, the general key recommendation of considering insights from SETI research to achieve more rigour in the assessment of algorithmic general intelligence claims has been remarked in an important contribution by Begoli and Sadovnik [17]. In another context, Vallor [60] utilizes the metaphor of present-day "AI" as a mirror for humanity the passive contemplation and reification of which risks to hold back humanity from reinventing an unknown open future. Given the challenges of the deepfake era, it seems straightforward that humanity has the choice to *actively* harness the scientific evaluation of algorithmic superintelligence claims as a SETI-inspired cosmic mirror for purposes of self-comprehension in order to achieve the urgently needed future amelioration of epistemic resiliency and algorithm-related literacy [62]. As suggested by Vallor, the future also calls for crucial technomoral virtues [58–60] facilitating an improved meaningfulness [57].

On the whole, as long as there exists *a* scientifically analyzable task of interest *to* humanity as a whole that human civilization *can* fulfil with higher reliability than an algorithm y (be it *currently* of economical value or not, a property that is habitually fluid and subject to change), it would be *scientifically* negligent and irrational to declare that y is absolutely superintelligent or that y is superintelligent in relation to the entire human civilization in all tasks of interest to humanity and would have made human civilization obsolete. Even from a misanthropic greedy economic perspective, as long as there exists a scientifically analyzable task of interest to humanity which humanity can solve with arbitrary higher reliability than an algorithm, it is conceivable that this could impact the economical value of that task – by what humanity stays relevant to itself. In this context, the subtle implications of the relativity of instantiated intelligence must be noted. While some may argue that one could shift to claim algorithmic superintelligence in specific tasks, this is a dubious move as even simple already existing tools including calculators would then be declared to be superintelligent in a specific task – a practice that would establish the meaninglessness of a grandiose term that currently bootstraps immense fears and projections. In short, we recommend a *deflationary* utilization of the term algorithmic superintelligence and provide a clear definition in Section 3. In general, it is obvious that many algorithms already surpass humanity in tasks of epistemic matter[3] (EM) repeating *within a given paradigm.* The latter makes algorithms better EM repeaters within a paradigm. Moreover, algorithms can perform epistemic dark matter (EDM) mining and epistemic dark energy (EDE) generation with arbitrary lower latency than humanity within a given paradigm which makes them lower-latency EDM miners and lower-latency EDE generators than humanity within the boundaries of a given paradigm. However, because of an emerging trade-off between lower latency and higher reliability, algorithms are *not* necessarily higher-reliability EDM miners and higher-reliability EDE generators in comparison to humanity (see e.g. [16, 37, 39]). Finally, human civilization is capable of generating epistemic tunneling (ET) events of universal scope with arbitrary high reliability but fundamentally unpredictable latency. The generation of the inherently *explanatory* (as relevant in critical rationalism [50, 49], its recent regimentation [29, 28] and adaptation to the idiosyncracies of the deepfake era [3, 5]) universal ET events has been described as a non-algorithmic task [6].

2 A Theoretical Solution

In our view, the most rigorous scientific evaluation frameworks of algorithmic *superintelligence* claims[4] (which subsume a claim of algorithmic general intelligence) in present-day humanity would at least necessarily require: 1) taking the

[3] As proposed in [6], we metaphorically compartmentalize the “epistemic cosmos" as follows: both the known known (i.e. the currently best theories expressible as so-called explanatory blockchains (EBs)) and the known unknown (i.e. open questions) form what one can term epistemic matter (EM), the unknown known (i.e. new but non-EB-like information that is consistent with EM but yet hidden) is referred to as epistemic dark matter (EDM) while the locally unknown unknown (i.e. new non-EB-like information that is inconsistent with EM) is called epistemic dark energy (EDE). Beyond EDE, the currently locally inaccessible new better scientific and philosophical paradigms of the future are metaphorically described to be fundamentally unpredictably but yet one day achievable via what one can term epistemic tunneling (ET). Each ET event is paradigm-shifting and instantiates a novel previously inconceivable epistemic cosmos with new EM, new EDM and new EDE.

[4] A match of intelligence can obviously not be deduced from EM repeating. But neither can it be deduced from EDM mining nor EDE generation since those are also based on already available EM from a civilization like present-day humanity. The remaining valid task would be ET but the latter is inherently *transformative* and would precisely entail the EB-measurement of a *difference* in intelligence – by what one can only analyze claims of algorithmic *super*intelligence *in relation* to present-day humanity. "Human-level" "AI" is not a useful scientific expression.

entire current human civilization as a baseline (this signifies that one is evaluating algorithmic supremacy claims against an entire civilization and it implies that *all* hereto willing humans can play the role of evaluators), 2) crafting a new better explanatory blockchain [6] (EB) that explains *transparently* how the purported candidate algorithmic superintelligence has been built and why intelligence/creativity/consciousness would be algorithmic (i.e. provisionally refuting all old EBs that conjecture its non-algorithmic [23, 52] nature) *and* additionally 3) generating *multiple* successive *civilization*-level ET events inspired by scales from SETI research [6]. Multiple tests for what is termed "human-level" "AI" have been reported: The Turing Test [30], The Robot College Student [44], The Employment Test [43], The Ikea Test [64], The Coffee Test [64] and The Modern Turing Test [46]. Unfortunately, none of those interesting and thought provoking approaches but tailored to only subsets of humanity fulfils the mentioned three criteria. Beyond that, a few companies developed their own customized "AGI" definitions. For instance, one "Super AI" scheme comprises five levels [20]: 1) chatbots, 2) reasoners, 3) agents, 4) innovators and finally 5) organizations (i.e. an algorithm able to do the work of an organization). It is easily cognizable that this scheme does again *not* fulfil the three criteria. Note that the term reasoning is superfluous if it is simply describable as EDM mining. Strikingly, ET events have nothing in common with what is called "reasoning". They are fundamentally unpredictable creative leaps. Since algorithmic ET events are impossible, Level 4 is misleading. Since there is no algorithmic epistemic agency [18], Level 3 is ill-conceived too. Another taxonomy [32] from a different company proposes six consecutive levels from 0) no AI over 1) emerging ("equal to or somewhat better than an unskilled human") to 5) Superhuman ("outperforms 100% of humans"). Interestingly, it is claimed that Level 1 in the taxonomy has already been achieved. The latter reflects the epistemic biases mentioned in the last section. There is already now an unnecessary dehumanization and deanimization taking place in the definition of Level 1. Since there are no transparently specified tasks and evaluation procedures, only Level 5 fulfils but *one of the three* criteria. A simple but meticulous exemplary procedure for a scientific evaluation of algorithmic superintelligence claims that fulfils all three criteria and features three exemplary civilization-level ET events has been formulated recently [6].

3 Practical Implications of Theoretical Solution

As summarized in Appendix A, according to the currently best available EBs, it is both *impossible* for a civilization x to: 1) reliably build an algorithmic entity y that would be EB-measured to be superintelligent in relation to that civilization x and 2) reliably control a pre-existing entity w that would be EB-measured to be superintelligent in relation to x. Bizarrely, there is an idea circulating in the present information ecosystem that definitions for algorithmic general intelligence (so-called "AGI") are a matter of personal taste or can act as brands for companies. The latter is equivalent to asserting that the new better provisional scientific definition of what should count as a perpetual motion machine in thermodynamics should be decided on the basis of a social-media-enacted democratic poll incorporating arbitrary subjective preferences such as claims of "feeling the perpetual motion in the sky" as promoted by the demo of a random company. Against this backdrop, here is a clearer and more transparent scientific definition of algorithmic superintelligence (subsuming the property of algorithmic general intelligence): relative to present-day humanity, an algorithmic superintelligence would be an algorithm able to generate *arbitrary many* successive *civilization*-level ET-based tasks of interests to current humanity with arbitrary higher reliability and *arbitrary lower latency* than the entire present-day human civilization could. In short, an algorithmic superintelligence would either be a proper subset of or de facto be a so-called *epistemic perpetuum mobile* [11]. Firstly, while the latter is impossible, already claiming to be able to implement it represents a severe epistemic security risk for humanity linked to the π-Doom scenario [7]. On the one hand, an over-reliance on algorithms in *safety-critical* contexts risks to unnecessarily engender existential risks due to the *absence* of EB understanding by algorithms (a conundrum which will become increasingly relevant in the deployment of algorithmic so-called "agents" – another example of an oxymoron). On the other hand, superfluous fears of algorithms sustain the form of epistemic stasis that life cannot permit itself to undergo for an arbitrarily long time especially in the presence of malicious human actors who could exploit those algorithm-related epistemic vulnerabilities in humanity. Secondly, when not only stating that one is able to build an epistemic perpetuum mobile but also additionally insisting that one is able to control the latter, one professes to perform a double magical act beyond planetary scope. By way of simplified comical illustration, consider the following question: how could a civilization having a certain limited power production (say 10^{13} W) ever possibly control a universe-spanning entity of which it is a part of that could harness say 10^{46} W? Asking humanity to believe that superintelligence (which cannot even be built by a less-intelligent entity) is also controllable is asking humanity to reenter the dark ages where magical chemical shortcuts to gold production where sought after. While gold isotopes can be obtained from bismuth [2] using a technique in nuclear reactors, a reliable transmutation faces prohibitive costs and resources and is thus impossible. In analogy to no-free lunch theorems [1], there is no algorithmic shortcut to higher civilization levels. Epistemic jumps occurs via fundamentally unpredictable *non*-algorithmic civilization-level ET events.

4 Conclusion

This short commentary focused on different approaches for scientific evaluation frameworks given algorithmic superintelligence achievement claims in the deepfake era. Next to listing examples of such already existing evaluation frameworks, the paper elaborated on their suitability, discussed shortcomings and possible solutions. The paper explained why the term "human-level" algorithm is not a scientifically useful description and provided a clearer more transparent definition of algorithmic superintelligence (subsuming algorithmic general intelligence): relative to present-day humanity, an algorithmic superintelligence would be an algorithm able to generate *arbitrary many* successive *civilization*-level epistemic tunnelling (ET) tasks of interest to current humanity with arbitrary higher reliability and *arbitrary lower latency* than the entire present-day human civilization could (see also [6] for a simplified illustration). In short, an algorithmic superintelligence would either be a proper subset of or de facto be a so-called *epistemic perpetuum mobile* [11]. In public discourse, misguidedly *instating* assertions on the feasibility and inevitability of the epistemic perpetuum mobile, a form of *algorithmic chrysopoeia* practicability claim, could lead to multiple conceivable superfluous undesirable outcomes [27, 41, 45, 63] connected to an even larger epistemic attack surface that malicious actors could exploit effortlessly and to catastrophic risks via dangerous over-reliance on isolated algorithmic loops in safety-critical contexts. However, in addition to criticizing these concerning developments, it may be time to also consciously and transparently harness the underlying conundrum as a SETI-inspired [19] cosmic mirror for the entire human civilization in order to scientifically build up a higher-quality epistemic resiliency for the future via achieving an improved comprehension of intelligence/creativity/consciousness. The latter could reflect humanity's now *active* use of the algorithmic mirror [60] – for augmentative purposes. Via the focus on the scientifically analyzable SETI-related tasks of interest, such an *open science* endeavor could even be suitable for parties with disagreements. For those convinced of the absolute algorithmicity of humanity's creativity including their own, it would become possible to scientifically explore the topic unpretentiously without enacting the current tendency to spread utopia promises or doom prophecies on algorithmic supremacy *before* ever having provisionally refuted the *non*-algorithmiticity of human civilization via presenting a new better explanatory theory of consciousness/creativity/intelligence which also transparently elucidates how the algorithm has been built and additionally demonstrating *multiple* successive algorithmically-generated *civilization*-level ET tasks (e.g. up to an immediately actionable new better explanation on how to physically craft a new universe in a laboratory – a physically in principle possible [15, 25, 42, 55] but elusive ET task of interest to at least a proper subset of humanity which only a civilization that is much more advanced than present-day humanity, i.e. a civilization that would appear to be an "alien superintelligence" in relation to current humanity could satisfyingly solve). For humans who intend to investigate and unfold the *non*-algorithmic facets of human civilization, the project would entail a focus on how to epistemically elevate humanity as a whole by *stimulating* (instead of fruitlessly trying to algorithmically extinguish) human creativity in novel disruptive [22], transformative [21] ways. Oddly, it seems that precisely such semi-adversarial scientific collaborations within an open science framework could vivify humanity's creativity in science and philosophy. If humans were algorithms, there would be nothing to loose as this verdict would already apply to both present and past. If not, there is *at least* a (re)new(ed) life to win for those who engage in self-zombification. Perhaps, the mere consideration of the SETI-inspired algorithmic cosmic mirror could act as mental nourishment.

References

1. Adam, S.P., Alexandropoulos, S.A.N., Pardalos, P.M., Vrahatis, M.N.: No free lunch theorem: A review. Approximation and optimization: Algorithms, complexity and applications pp. 57–82 (2019)
2. Aleklett, K., Morrissey, D.J., Loveland, W., McGaughey, P.L., Seaborg, G.T.: Energy dependence of ^{209}Bi fragmentation in relativistic nuclear collisions. Phys. Rev. C **23**, 1044–1046 (Mar 1981). https://doi.org/10.1103/PhysRevC.23.1044, https://link.aps.org/doi/10.1103/PhysRevC.23.1044
3. Aliman, N.M.: Cyborgnetics – The Type I vs. Type II Split. Aliman, Nadisha-Marie (2021)
4. Aliman, N.M.: Cyborgnetic Invariance. Aliman, Nadisha-Marie (2023)
5. Aliman, N.M.: Epistemic Security Augmentation (2023)
6. Aliman, N.M.: Acentric Intelligence. PhilPapers (2024)
7. Aliman, N.M.: Epistemic Doom In The Deepfake Era. PhilPapers (2024)
8. Aliman, N.M.: Responsible AI Control. In: Trustworthiness and Responsibility in AI – Causality, Learning, and Verification. Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2024)
9. Aliman, N.M., Kester, L.: Facing Immersive "Post-Truth" in AIVR? Philosophies **5**(4), 45 (2020)
10. Aliman, N.M., Kester, L.: Epistemic Defenses against Scientific and Empirical Adversarial AI Attacks. In: CEUR Workshop Proceedings. vol. 2916. CEUR WS (2021)
11. Aliman, N.M., Kester, L.: VR, Deepfakes and Epistemic Security. In: 2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp. 93–98. IEEE (2022)
12. Aliman, N.M., Kester, L.: AI-Related Misdirection Awareness in AIVR. PhilPapers (2023)
13. Aliman, N.M., Kester, L., Wernaart, B.: Moral Programming: Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies, pp. 63–80. Wageningen Academic Publishers (2022)
14. Aliman, N.M., Kester, L., Werkhoven, P.J.: XR for Augmented Utilitarianism. 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) pp. 283–285 (2019)
15. Ansoldi, S., Guendelman, E.I.: Child universes in the laboratory. arXiv preprint gr-qc/0611034 (2006)
16. Basmov, V., Goldberg, Y., Tsarfaty, R.: LLMs' Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements. arXiv preprint arXiv:2404.06283 (2024)
17. Begoli, E., Sadovnik, A.: What Can AI Researchers Learn From Alien Hunters? . https://spectrum.ieee.org/artificial-general-intelligence-2668132497 (2024), IEEE; accessed 19-September-2024
18. Birhane, A., McGann, M.: Large models of what? Mistaking engineering achievements for human linguistic agency. Language Sciences **106**, 101672 (2024)
19. Charbonneau, R.: SETI, artificial intelligence, and existential projection. Physics Today **77**(2), 36–42 (2024)
20. Cook, J.: OpenAI's 5 Levels Of 'Super AI' . https://www.forbes.com/sites/jodiecook/2024/07/16/openais-5-levels-of-super-ai-agi-to-outperform-human-capability/ (2024), Forbes; accessed 05-September-2024

21. 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)
22. Cropley, D., Cropley, A.: Creativity and the Cyber Shock: The ultimate paradox. The Journal of Creative Behavior **57**(4), 485–487 (2023)
23. Faggin, F.: Possibilities are quantum. Possibility Studies & Society **1**(1-2), 67–72 (2023)
24. Fallis, D.: The epistemic threat of deepfakes. Philosophy & Technology **34**(4), 623–643 (2021)
25. Farhi, E., Guth, A.H., Guven, J.: Is it possible to create a universe in the laboratory by quantum tunneling? Nuclear Physics B **339**(2), 417–490 (1990)
26. Floridi, L.: Artificial intelligence, deepfakes and a future of ectypes. Philosophy & Technology **31**(3), 317–321 (2018)
27. Floridi, L.: Why the AI Hype is another Tech Bubble. SSRN:4960826 (2024)
28. Frederick, D.: Against the Philosophical Tide: Essays in Popperian Critical Rationalism. Critias Publishing (2020)
29. Frederick, D., et al.: Falsificationism and the Pragmatic Problem of Induction. Organon F **27**(4), 494–503 (2020)
30. Gonçalves, B.: The Turing test is a thought experiment. Minds and Machines **33**(1), 1–31 (2023)
31. Gros, C., Kester, L., Werkhoven, P., Martens, M.: Defining a method for ethical decision-making for automated vehicles. Researchgate (2023)
32. Heaven, W.D.: Google DeepMind wants to define what counts as artificial general intelligence . https://www.technologyreview.com/2023/11/16/1083498/google-deepmind-what-is-artificial-general-intelligence-agi/ (2023), Forbes; accessed 05-September-2024
33. Holbrook, C., Holman, D., Clingo, J., Wagner, A.R.: Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. Scientific Reports **14**(1), 19751 (2024)
34. Horvitz, E.: On the Horizon: Interactive and Compositional Deepfakes. arXiv preprint arXiv:2209.01714 (2022)
35. 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)
36. Kalpokas, I., Kalpokiene, J.: On alarmism: between infodemic and epistemic anarchy. In: Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation, pp. 41–53. Springer (2022)
37. Katz, M., Kokel, H., Srinivas, K., Sohrabi, S.: Planning with Language Models Through The Lens of Efficiency. arXiv preprint arXiv:2404.11833 (2024)
38. Kozinets, R.V., Gershoff, A.D., White, T.B.: Introduction to special issue: trust in doubt: consuming in a post-truth world. Journal of the Association for Consumer Research **5**(2), 130–136 (2020)
39. Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H.T., Gurevych, I.: Are Emergent Abilities in Large Language Models just In-Context Learning? arXiv preprint arXiv:2309.01809 (2023)
40. 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)
41. Markelius, A., Wright, C., Kuiper, J., Delille, N., Kuo, Y.T.: The mechanisms of AI hype and its planetary and social costs. AI and Ethics pp. 1–16 (2024)
42. Merali, Z.: A big bang in a little room: the quest to create new universes. Hachette UK (2017)

43. Mikhaylovskiy, N.: How do you test the strength of AI? In: International Conference on Artificial General Intelligence. pp. 257–266. Springer (2020)
44. Mondol, M.A.R., Pothula, A., Park, D.: A definition and a test for human-level artificial intelligence. arXiv preprint arXiv:2011.09410 (2020)
45. Monett, D.: Sabbatical Research Project "AI promises, fallacies, and pitfalls: Inhibitors and stepping-stones for progress in Artificial Intelligence". Researchgate (2024)
46. Morris, D.: Magical thinking and the test of humanity: we have seen the danger of AI and it is us. AI & SOCIETY pp. 1–3 (2023)
47. Nagy, P., Neff, G.: Conjuring algorithms: Understanding the tech industry as stage magicians. new media & society **26**(9), 4938–4954 (2024)
48. Placani, A.: Anthropomorphism in AI: hype and fallacy. AI and Ethics pp. 1–8 (2024)
49. Popper, K.: The logic of scientific discovery. Routledge (1959)
50. Popper, K.: Conjectures and refutations: The growth of scientific knowledge. Routledge (1962)
51. Reed, N., Leiman, T., Palade, P., Martens, M., Kester, L.: Ethics of automated vehicles: breaking traffic rules for road safety. Ethics and Information Technology **23**(4), 777–789 (2021)
52. 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)
53. Rosenberg, L.: Prepare for arrival: Tech pioneer warns of alien invasion . https://venturebeat.com/ai/have-we-reached-peak-human/ (2022), VentureBeat; accessed 19-September-2024
54. Rosenberg, L.: Have we reached peak human? . https://venturebeat.com/ai/have-we-reached-peak-human/ (2024), VentureBeat; accessed 19-September-2024
55. Shainline, J.M.: Does cosmological evolution select for technology? New Journal of Physics **22**(7), 073064 (2020)
56. Spatola, N., Urbanska, K.: God-like robots: the semantic overlap between representation of divine and artificial entities. Ai & Society **35**(2), 329–341 (2020)
57. Steen, M.: Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Journal of Moral Philosophy **18**(1), 87–90 (2021)
58. Vallor, S.: Technology and the virtues: A response to my critics (2018)
59. Vallor, S.: Twenty-first-century virtue. Science, technology, and virtues: Contemporary perspectives p. 77 (2021)
60. Vallor, S.: The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking. Oxford University Press (2024)
61. Verspoor, K.: A new 'AI scientist' can write science papers without any human input. Here's why that's a problem. https://theconversation.com/a-new-ai-scientist-can-write-science-papers-without-any-human-input-heres-why-thats-a-problem-237029 (2024), The Conversation; accessed 24-August-2024
62. Vrabič Dežman, D.: Promising the future, encoding the past: AI hype and public media imagery. AI and Ethics pp. 1–14 (2024)
63. Westerstrand, S., Westerstrand, R., Koskinen, J.: Talking existential risk into being: a Habermasian critical discourse perspective to AI hype. AI and Ethics pp. 1–14 (2024)
64. Wikipedia: AGI. https://en.wikipedia.org/wiki/Artificial_general_intelligence (2024), Online; accessed 05-September-2024

A Cyborgnetic Invariance

Cyborgnetic invariance [6] implies that intelligence/creativity/consciousness is relative in all cases except in the case of the invariantly maximal quantity superintelligence level α which is crucially *non*-algorithmic. While instantiated intelligence is relative, the invariantly maximal level α is *generic* and does *not* have an own frame of reference within the universe (i.e. it can never be fully instantiated in matter). In analogy with the acentric model of the expanding cosmos, it introduces the acentric notion of intelligence, creativity and consciousness. It also entails that: 1) a quality superintelligence is impossible, 2) it is impossible for x to reliably build and entity y that is EB-measured to be superintelligent in relation to x – i.e. that a) to reliably build a quantity algorithmic superintelligence is impossible and b) a narrow algorithm recursively self-improving to a general intelligence is impossible, and 3) it is impossible for x to reliably control an entity y that would be EB-measured to be superintelligent in relation to x. It is a scientific theory amenable to experimental problematization in present-day human civilization via letting the purported algorithmic superintelligence candidate – irrespective of the specific algorithmic paradigm – generate multiple successive civilization-level epistemic tunneling (ET) events via a procedure specified earlier [6] and it is formulated such that it could be provisionally refuted by additionally providing a new better EB on how the algorithm has been built and why intelligence/creativity/consciousness would instead be algorithmic. Cyborgnetic invariance does *not* imply the impossibility of building *a* general intelligence. It does however entail the impossibility of an *algorithmic* general intelligence and the impossibility of a civilization D building an entity C that would be EB-measured to be superintelligent in relation to D. In this paradigm, it *is* possible that via an unpredictable ET event, a civilization A builds a *non*-algorithmic entity D "from scratch" that could subsequently, at an *unpredictable* future time point, decide to corroborate its new situation as a *non*-algorithmic general intelligence (NaGI), an EB-transformed civilization C. From this EB-measurement, civilization C could conclude that it now became superintelligent in relation to the civilization D it once was. In short, it is in theory possible for civilizations that are much more advanced than present-day humanity to indirectly build a NaGI "from scratch" via an unpredictable ET event. However, *multiple* steps separate present-day humanity from that, so it is currently no imminent topic. Concerning the theoretical option for A of co-creating D starting from seemingly suitable pre-existing non-algorithmic biological entities as NaGI strategy, there is no guarantee on when or if their potential future transformed civilization C would choose to corroborate their own general intelligence. Hence, on the whole, a sincere present-day NaGI project would first mean a focus on humanity's self-comprehension.

B The Conjecture, Observe, Orient, Co-create, Act Loop

Algorithms are neither "agents" [35] nor are they "general purpose" entities and should *not* be selected as standalone entities in safety-critical contexts. Instead, a local encapsulation of algorithms into people-controlled units (i.e. which locally insert people-crafted pluralistic moral models [13, 14, 31, 51]) is needed. While it is often maintained that humans have to rely on algorithms to stay significant, the assumption is flawed since following the currently best EBs, it holds that no algorithmic support can ever guarantee that a civilization *will* instantiate a new universal ET event at a certain time. In safety-critical contexts, there is the need for the meta-paradigm of a so-called COOCA loop instead of insufficient algorithmic OODA-loops where on-the-fly adaptation via EB comprehension is impossible by-design. Some past approaches are already compatible with the generic COOCA-loop meta-paradigm and can be integrated as follows:

- ***Inter*-function-level:** There must be at least one person in each single function to anticipate for the eventual case of EB-based communication *between different functions*. This is instantiated by some human-in-the-loop approaches. Also, recall that an algorithm in a function is *not* obligatory.
- ***Intra*-function-level:** While each high-level function must contain at least one person, there is room for improvement *within* an individual function if needed. There, where feasible, one can improve the situation by harnessing small locally encapsulated algorithmic OODA loops. This allows any of the three paradigms locally enclosed *within* one unit containing at least one person: human-before-the-loop, unsupervised loop and human-in-the-loop.

C Algorithmic versus Non-Algorithmic

There have even been statements that so-called "AI" models with an interaction mode with people are non-algorithmic. However, such claims must be reassessed against the background of the relativity of instantiated intelligence. For example, a software x extended by real-time interactions with present-day human civilization can be interpreted to be non-algorithmic *in relation* to another *algorithm* y without real-time interactions with present humanity. However, in relation to present-day human civilization, the software x would still stay algorithmic. Present-day humanity is neither forced nor pre-determined to keep interacting with x. In sum, in relation to the entire current human civilization, all software built *by* human civilization is algorithmic. It seems bizarre and untenable to artificially attempt to enforce the perspective of an algorithm on human civilization. Due to the relativity of instantiated intelligence/consciousness/creativity, it does hold that there *can* exist frames of reference from which a conscious civilization and an algorithm would *appear* indistinguishable. However, the core scientific statement is that consciousness fundamentally precedes algorithms because the only invariantly maximal superintelligence level is non-algorithmic as EB-measured from *all* frames of reference. Why the latter is a scientific statement amenable to experimental problematization has been explained elsewhere [6]. From the perspective of human civilization, all human-built software stays algorithmic while the *future* of life's evolution including other animals in the biosphere is *non*-algorithmic. Life is more complex than the software it creates. While it may be tempting to assume that it would suffice to couple a human-made software with a hardware that makes use of automated measurements of genuine physical randomness to produce high-quality non-algorithmic random outputs, it is important to note that such a software would still *not* be able to reliably (i.e. with arbitrary high accuracy) generate universal ET events. Random numbers are not sufficient to encode meaning. On the whole, the mentioned software would still appear mechanistic and algorithmic to humanity not only because one could publicly access the discrete random inputs based on which the software functions but especially also because one could already predict a priori that whatever it generates, it will never be able to produce universal ET events with *arbitrary higher reliability* and *arbitrary lower latency* than humanity as a whole could. The latter is again a scientific statement amenable to experimental problematization (see the civilization-level scientific evaluation framework illustrated in [6]). Having said that, the implementation of a such a randomness-driven chaotic software would not only be useless when it comes to EB creativity, but it would obviously represent a severe security risk in safety-critical contexts – by what it should never be applied as standalone in those contexts. The latter is due to a *lack* of EB understanding and *not* due to the magical emergence of any algorithmic superintelligence in relation to present-day humanity.