AI and Ethics
https://doi.org/10.1007/s43681-024-00464-z

ORIGINAL RESEARCH

Talking existential risk into being: a Habermasian critical discourse perspective to AI hype

Salla Westerstrand[1] · Rauli Westerstrand[2] · Jani Koskinen[1]

Received: 30 November 2023 / Accepted: 5 March 2024

Abstract
Recent developments in Artificial Intelligence (AI) have resulted in a hype around both opportunities and risks of these technologies. In this discussion, one argument in particular has gained increasing visibility and influence in various forums and positions of power, ranging from public to private sector organisations. It suggests that Artificial General Intelligence (AGI) that surpasses human intelligence is possible, if not inevitable, and which can—if not controlled—lead to human extinction (Existential Threat Argument, ETA). Using Jürgen Habermas's theory of communicative action and the validity claims of truth, truthfulness and rightness therein, we inspect the validity of this argument and its following ethical and societal implications. Our analysis shows that the ETA is problematic in terms of scientific validity, truthfulness, as well as normative validity. This risks directing AI development towards a strategic game driven by economic interests of the few rather than ethical AI that is good for all.

Keywords Artificial intelligence · Habermas · AI hype · Existential threat argument · AI ethics · Critical discourse theory

1 Introduction

Recent developments in Artificial Intelligence (AI) have sparked a range of discussions about the societal and ethical implications of AI systems. Several studies have disclosed negative impacts of AI systems on people's lives through, e.g., discrimination [1–4], environmental hazards [5, 6] and biometric mass surveillance [7, 8]. On the other hand, AI comes with a myriad of opportunities in different sectors, such as enhancing medical treatment and diagnostics [9, 10] and promoting political participation [2, 11]. Meanwhile, a group of scholars and tech leaders have argued that the recent developments in generative AI show signs of Artificial General Intelligence (AGI), or perhaps even the beginnings of superintelligence, which—if not controlled—could lead to human extinction [e.g., [12]]. This argument starts from the premise that developing an artificial agent that surpasses human intelligence is both possible and inevitable—hence creating an existential risk for humanity in the future. Let us call this the *Existential Threat Argument (ETA)*. Whereas notably scholars associated with the effective altruism movement have talked about such a threat for some time now [13], in the midst of the current AI hype, business leaders seem to be increasingly jumping in to spread the doomsday message of a looming existential threat.[1] Recently, this culminated in an Open letter of the Future of Life Institute (FLI) published in March, 2023,[2] which was signed by over a hundred high-profile AI professionals and business leaders calling for a pause in efforts to create AI models more powerful than GPT-4.[3]

Salla Westerstrand and Rauli Westerstrand have contributed equally to this work.

✉ Salla Westerstrand
salla.k.westerstrand@utu.fi

Rauli Westerstrand
rauli.westerstrand@outlook.com

Jani Koskinen
jasiko@utu.fi

[1] Information Systems Science, University of Turku, Rehtoripellonkatu 3, Turku 20500, Finland

[2] Center for Philosophy, Disruptive Futures Institute, 8033 Sunset, Boulevard #276, Los Angeles, CA 90046, USA

[1] See the statement "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" published on the website of Center for AI Safety: https://www.safe.ai/statement-on-ai-risk

[2] "Pause Giant AI Experiments: An Open Letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[3] GPT-4 is a multimodal large language model created by OpenAI.

Published online: 02 April 2024

We argue that relying on the ETA based on AGI is misleading. In this paper, we show the problematic nature of feeding AI hype with the ETA, which reveals itself when these arguments are approached as a *discourse*. From this perspective, when we discuss technology and present arguments about its nature, we construct meanings and thus also our understanding of the technology itself. Discourses such as the ETA shape the ways in which we develop, regulate, and use technologies. As Swanson and Ramiller [14] have shown already a couple of decades back, discourse around emerging technologies is a prerequisite for the spread of innovation and can thus be seen as constructive of technology. Therefore, what kind of discourse we produce affects the direction which our technology development is heading. Besides dictating the contents of the discourse, gaining control over who gets to talk, when, where and how is a form of accumulating social power over people's actions. As Van Dijk [15, p.9] notes, the one who controls the discourse has the power to influence people's "knowledge, opinions, attitudes, ideologies, as well as other personal or social representations", leading to an indirect control over people's actions. When such control is in the interest of the one controlling but against the interest of the one being controlled, we can talk about power abuse [15].

Therefore, when proposing an alleged existential threat caused by AGI, we influence the way our future technologies are being developed. This can be expected to come with societal and ethical implications. For instance, although raising concerns about the impacts of AI is justified and needed, relying on the ETA risks directing attention and funding away from existing issues regarding Artificial Narrow Intelligence (ANI) towards hypothetical risks of a hypothetical technology that does not exist. We argue that the mechanisms behind such implications are not necessarily technical but social, which requires an approach that extends beyond the mere technical features and taps into the socio-technical dimensions of AI systems.

Hence, to better understand the impacts of the ETA discourse on the current direction of AI development and its following implications, we examine it using Jürgen Habermas's critical theory of communicative action with a view of analysing its truth value, truthfulness and rightness (normative validity). Based on this analysis, we offer critique and recommendations for directing AI discourse towards justified concerns that merit our attention in terms of resources such as research and monetary investments.

This work offers a contribution in two ways. On the one hand, it increases the conceptual understanding of the impacts of how we discuss AI systems and the related risks and opportunities. We bring forth philosophically robust examples of how the ways in which we talk about technology impact the very development of these systems, as well as the societal conditions for doing so. On the other hand, it contributes to the critical discussion around the implications of the current direction of AI development as a broader societal phenomenon. In times where academic research on the implications of AI hype and related phenomena is still limited, we consider such a contribution to be especially valuable for ensuring the technologies we develop truly benefit the whole of our societies.

In what follows, we first give a detailed description and conceptualisation of the ETA and how it is approached as a discourse in the present paper. We then engage in an analysis of societal and ethical implications of the discourse, and end with conclusions and discussion.

2 Existential threat argument, effective altruism and longtermism in AI hype

Talking about existential threats posed by advanced technologies is not a recent phenomenon. Science fiction writers have for long created scenarios of machines more powerful than humans taking over the world, painting both utopian and dystopian pictures of such futures. In the current discussions, however, the proposed existential threats of AI are not being discussed as science fiction but as actual concerns that allegedly could be realised in the near future.

In this paper, the term Existential Threat Argument (ETA) refers to statements according to which *developing an AGI that surpasses human intelligence is possible, if not inevitable, and which can—if not controlled—lead to human extinction.* It is thus essential to recognise that in this paper, the ETA does not encompass all types of existential risks proposed as a result of AI development. We focus here on the ETA that is currently actively advanced by several tech leaders and AI developers holding significant power over the direction of AI development.

In discussions concerning AI systems, the ETA manifests itself as commentaries proposing that the current developments in AI could eventually lead to a machine that surpasses human intelligence, bringing forth concerns around human extinction. Some have gone as far as suggesting that AGI is already here [16]. The discussion has taken place on multiple forums: Several authors have contributed to it through popular books on AI [e.g., [17–19]], and in the aftermath of the launch of popular APIs for Large Language Models, such as ChatGPT, open letters calling for attention to long-term impacts have emerged,[4].[5] Some governmental agents have also adopted

[4] "Pause Giant AI Experiments: An Open Letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[5] See the statement "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pan-

the ETA in their national AI policy. For example, in the last year, several news outlets have reported comments given by British government representatives—notably the prime minister Rishi Sunak—that warn about the existential threats of AGI development [20]. Following meetings with the heads of companies including Google's DeepMind and OpenAI, the media has discussed how the Prime Minister Rishi Sunak seems to have adopted the corporate messaging on the existential threats of AI [21] and has been suggested to ground British AI politics in effective altruism [22]. The ETA message has been echoed by other policy-makers, such as the Deputy Prime Minister Oliver Dowden in his speech to the UN General Assembly in September 2023.[6] Even before the UK government's global AI summit took place on November 1–2 2023, the event received critique amongst AI experts over its emphasis on existential threats rather than ongoing issues [23]. This led to an open letter to the Prime Minister Rishi Sunak, published on October 30th, 2023 and signed by 122 NGOs,[7] calling out the dominance of industry voices in the public realm, as well as the neglecting of issues relevant for people represented by these organisations.

Many of the proponents of the ETA in the context of AI seem to share certain common ideas arising from effective altruism (EA) and longtermism. Understanding these ideas can provide us with valuable contextual knowledge when evaluating the mechanisms of implications. For the purposes of this analysis, a concise—although reflective—introduction will do. According to Schopmans [13], the ETA in AI has ties to EA, which is a movement and a philosophy that—in their own words—"aims to find the best ways to help others, and put them into practice".[8] Although seemingly innocuous, if not admirable, both the practices and the philosophy of the movement have received critique (see, e.g., the discussion of critiques in [24]). For example, the philosophical foundation of the movement stresses a "duty to use a proportion of one's resources" [25, p.xiii] on charitable pursuits. This is connected with the idea of making as much money as possible during one's career to then be able to donate as much as possible. Ioannidis [26, p.42] points out, however, that such reliance on a capitalist understanding of how the world should function leads to neglecting the societal aspects of poverty that cannot necessarily be solved in monetary terms. Although EA has been occasionally extended to means other than donating to charity—a unit of resource being, e.g., an hour of labour [27]—the accumulation of wealth still appears as a foundational element of the ideology [24].

To decide where the efforts—accumulated capital, labour, or other resources—should be directed, many EA proponents have relied on the idea of *longtermism*. Accordingly, instead of thinking how we can solve problems in the short term, we should concentrate our efforts on ensuring that everything will be good in the far future [28]. Strong longtermism goes one step further in suggesting that "the impact on the far future is the *most* important feature of our actions today" [28, p. 2]. Here, longtermism has rightly received critique regarding our poor ability to predict the future, as well as the epistemic challenge of distinguishing actions that hold potential to improve the future from actions that do not [29]. Even so, longtermism has been doggedly persistent in discussion regarding AI, taking the form of concerns around existential threats posed by a supposedly intelligent machine. Part of the longtermist account is the idea that sentient AI is indeed possible, and that, consequently, the concern about the governance of superintelligence should be considered among the most pressing current issues [28], over and above more immediate AI-related issues.

It must be noted that neither EA nor longtermism are robust, established ideologies with solid philosophical foundations. Although there are some shared resources crafted by early proponents of the movements [e.g., [25, 28]], its philosophical foundations are still rather weak [e.g., [26]]. There are no lists of people nor are there many who would have publicly announced an affiliation with the movement. Moreover, being a relatively recent movement advanced by a small number of people, no significant body of scientific knowledge exists regarding either of the two ideologies. Therefore, it is likely that there are ETA proponents who do not identify themselves as part of either of these movements. Understanding this ideological background is, however, an important element when evaluating the ETA from the perspective of validity claims, which we next discuss in detail.

3 Critique of ETA-based AI Hype: validity claims

To conceptualise and critically evaluate the ETA and its implications, we have adopted a critical discourse perspective based on Jürgen Habermas's critical theory of communicative action. Habermas built his theory on the German tradition of critical theory, which can be traced back to the Frankfurt School philosophers,

Footnote 5 (continued)

demics and nuclear war" published on the website of Center for AI Safety: https://www.safe.ai/statement-on-ai-risk.

[6] Deputy Prime Minister Oliver Dowden's speech to the UN General Assembly on 22 September 2023 on the UK Government website: https://www.gov.uk/government/speeches/deputy-prime-minister-oliver-dowdens-speech-to-the-un-general-assembly-22-september-2023.

[7] https://ai-summit-open-letter.info/.

[8] https://www.effectivealtruism.org/articles/introduction-to-effective-altruism.

such as Max Horkheimer, Theodor Adorno and Herbert Marcuse [30, 31]. According to Habermas, reality is "objectified" in intersubjective communication [32, p.8], i.e., communicative rationality. In an ideal situation, it is based on communicative action, which aims towards mutual understanding [33, p.286]. In order for a speech act to contribute to communicative action, it needs to fulfil three validity claims: rightness, truthfulness, and truth. By *rightness*, or legitimacy, Habermas refers to action that follows normative rules of the given normative context [33, p.306]; by *truthfulness* that the speaker "means what he says", thus linking truthfulness to the subjective experience or intention of the speaker which should be communicated truthfully to the person in the receiving end [33, p.307]; and by *truth* that the underlying presuppositions in the utterance are in accordance with the existing, intersubjectively shared context [33, p. 306]. Accordingly, a speech act can always be rejected as strategic rather than communicative if one of these conditions is left unfulfilled.

According to Habermas, only communicative action that applies validity-based arguments can provide a procedural basis for moral judgements that are critically contested and therefore acceptable as a basis for Discourse ethics [34]. Discourse ethics focuses on discourse as a practice where the goal is to reach an agreement or a common will that is based on a deliberative and rational argumentation process between all parties that are concerned. In his book *Between Facts and Norms* [35], Habermas also underlines deliberation, rationality and exclusion of strategic games as essential to ethical discourse. Strategic games are ways of influencing other participants in discourse by means other than providing a better argument. Bargaining, hidden agendas, and use of authority over others are some examples of such strategic actions [36].

In the context of the ETA, we thus approach the argument as the speaker's contribution to shared communication that influences how future technologies are developed, and thus what kind of societal or ethical implications it is likely to have. Are proponents of the ETA contributing to communicative action, or perhaps playing a strategic game? To assess this, and to gain understanding of the underlying assumptions of the arguments, we apply Habermas's validity claims to the ETA. First, we examine the truth, i.e., scientific validity of the argument to see if it is factually correct in the light of available knowledge. By addressing truthfulness, we can evaluate if the speaker means what they say and thus communicates sincere beliefs to the audience. Finally, by reflecting on the rightness of the argument, we can reveal its connections to the prevailing normative rules and societal structures.

3.1 Truth: the nature of artificial intelligence

The first step is to analyse whether the argument posed by ETA proponents can be accepted with regards to the validity claim of *truth*, i.e., whether the speaker makes a true argument, based on "correct existential presuppositions" [33, p. 307]. The ETA as defined in this paper (see Section 2) is based on the presupposition that developing an AGI, or superintelligence—which would then become an existential threat emerging from the technology itself—is possible, even likely in the near future. This presupposition is, however, subject to debate.

The idea of building an intelligent artefact comes from a commitment to scientific physical reductionism. Scientific reductionism, in turn, follows from a commitment to a philosophical position (material monism) according to which it must be possible to explain everything in terms of material constitution (physics). Proponents of this position hold that physical phenomena are fundamental in all nature, and that, consequently, all things, including self-consciousness and thus intelligence, too, result as a by-product of some ground-level physical phenomena (epiphenomenalism) [37].

If we buy into the idea of material reduction, then it easily seems plausible that what we describe as self-consciousness and intelligence emerge in our material brain and are therefore reducible to material. If so, then it seems plausible to suggest that if we can through physical sciences discover how material interaction generates self-consciousness and intelligence, we will be able to build artefacts that manipulate matter in an architecture similar to the laws and rules that govern the brain. If we achieve this, we will have built an intelligent machine. Such ideas are expressed by many philosophers [e.g., [38, 39]], neuroscientists [e.g., [40–43]], and AI researchers [e.g., [44, 45]], but here we will refer to AI researchers Stuart Russell and Peter Norvig [46], who list the reductionist question: 'How does the mind arise from a physical brain? (p. 24). This is the foundational idea behind the argument that AGI could be possible.

This reductionist question is, however, problematic: we are dealing with a conceptual confusion that appears to make sense, but upon closer examination is illogical. While the brain is a causal necessity without which self-consciousness and intelligence of a creature would not exist [e.g., [47]], there is an extensive body of knowledge in the field of philosophy of mind that suggests that consciousness/intelligence are not properties attributable to the material of the brain. Instead, consciousness and intelligence are *concepts* that exist not merely as a result of a material construction but because of, e.g., meanings attributed to them—meanings that are not materially composed. As Bennett and Hacker [48] put it,

> "Legal systems consist of laws and not of matter; poems consist of stanzas, not of ink; and revolutions consist of human action and events. The materialist might grant that this is what laws, and poems, and revolutions *consist* of, but deny that they are *made* of anything. We can concede this too. But even if it is true that everything that is made of anything is made of matter, this thesis goes no way to sustain any form of ontological reduction according to which all 'entities' are reducible to material entities. Nor does it support any form of explanatory reduction according to which the properties and behaviour of everything that exists are to be explained in terms of the properties and behaviour of its constituent matter." (p. 359)

Yet, according to some eminent neuroscientists [e.g., [40–43]], it is the material brains instead of humans as entities that see and hear things, build models, know things, reason inductively and present arguments, among other things. Problematically, as the subject matter is difficult, such statements often go unchallenged.

The root cause of such confusion is understandable: the explanatory power of science is limited by the fact that science can only concern the mechanics of the things it studies while it must presuppose as given the things that it means to study: otherwise, science would lack an object of study upon which it makes its claim of objectivity. Hacker and Bennett [48] identify a transgression of the explanatory power of science by what they call the mereological fallacy. Falling for the mereological fallacy means believing that an explanation for a thing can be given by examining the constituents of the thing and giving an account of how those constituents work together. Believing in such an explanation is the proposed solution of reductionism, specifically manifest in the concept of *emergence*: the idea that higher level properties emerge from the interaction of underlying constituent parts.

The fallacy involved in a mereological explanation was already identified by Descartes, who warned that in order to explain vision we must avoid giving an explanation that assumes an extra pair of eyes in the brain that see the image that is projected through the eyes. Despite this, Descartes himself proceeded against his own better judgement and went on to propose that the soul can look at images that are projected through the eyes. This fallacy is constitutive of Cartesian dualism.

Reductionists recognise this weakness involved in a mereological argument and therefore attempt to refine their explanation. They try to avoid presupposing vision by suggesting it *emerges* from the matter constituting the brain. The reductionist proposal is that the eyes receive symbolic information, which is then supposed to effect neurons in such a way that the neurons receive the information contained in the symbols and carry it forth for further decoding by some reader (the brain) that converts the symbols into an image which is then seen.

But such an explanation merely repeats Descartes' fallacy. The only difference is that instead of the soul seeing an image coming through the eyes onto the pineal gland, as Descartes held, some scientists now believe that the brain constructs representations from data or pieces of information carried through the eyes in the form of light arrays. Just the same as in the case of Descartes' proposition, these scientists fail to explain what they sought out to explain and have now involved themselves in a dualist form of explanation. This is a fundamental epistemological error of transgressing the meaning that is given to the words that enable the articulation of scientific thought itself. Although not in the scope of this paper, it is worth noting that once this boundary of meaning—of making sense—has been transgressed, the ideas that reductionists hold would lose their scientific validity.

On the other hand, phenomenological philosophy has revealed the problems of cognitivism, i.e., the view based on the hypothesis that the mind is a computational system [49]. Phenomenological philosophies offered by Husserl [50], Heidegger [51] and Merleau-Ponty [52] indicate that our cognition is not mere material computation ability that can be replaced by computers. As such, there is little hope that AI could be generated through a top-down, formalist approach [49]. This is in line with Habermas's view [33, 53] that society consists of systems and lifeworlds, both of which are characterised by distinct rationalities. According to Habermas, lifeworld is where our lives are created and shared. Lifeworld is a dimension of everyday life of humans, which dimension is shared by individuals connecting with the lifeworlds of one another. Systems, on the other hand, refer to economic, political, technological and administrative systems, where actions serve the institutionalised goals of the systems that easily colonise the lifeworld. Like Fairtlough [54] stated, our ability to understand the world and ourselves arises from our communication with each other. Accordingly, consciousness arises not from the brains, but rather from there existing an embodied being in the circumstances of life.

It is essential to note here that the role of language in consciousness/intelligence is not reduced to the capability of producing correct syntax with a sufficient level of semantic coherence. As discussed by several scholars [e.g., [5, 55–57]], language technologies, such as Large Language Models (LLMs) are based on statistical calculations of the probabilities of certain words appearing in the same sentence. It does not require—and hence does not imply—understanding or knowledge of the context. Language produced with LLMs arises from the motivation of the programmer to produce as human-like sentences as possible

with a machine. By contrast, in conscious communication by humans, producing coherent language is not the end goal but a vehicle for fulfilling an action driven by our motivations, which contributes to the creation of shared understanding of meanings and concepts. What is essential here is the ability to proactively communicate with others in a society—the phenomenology of being, where mind is bounded to our body but our lives are connected with other people by social practices [see e.g [51]].

It thus seems unreasonable, to say the least, to root the future of AI-driven societies to the premise that conceptual phenomena could be given a physical explanation in terms of particles, fields, and waves. Try to make sense of providing a physical explanation, for example, of the 2008 financial crisis: did this crisis happen on an atomistic level which then brought about the crisis on a higher-level?

Russell and Norvig [46, p.19] present the definition of intelligence taken on in AI research as 'doing the "right thing" ' in the right circumstances: so, exhibited in plausible behaviour. But this idea comes from control theory and cybernetics, and an often-cited example is that of the thermostat. Russell and Norvig aver that such "invention[s] changed the definition of what an artefact could do. Previously, only living things could modify their behaviour in response to changes in the environment" (p.33). However, such examples fall into the trap of anthropomorphising things, which in the context of AI has been shown problematic, if not dangerous [e.g., [58]], blurring the scientific validity of the statements presupposing the possibility of superintelligent machines. It seems that AI is more of a marketing term that makes the tech sound more interesting and advanced,[9] or a subject for science fiction [48, p. 299].

Research and the following debate around the nature of intelligence and consciousness is, of course, much broader than what we are able to discuss in the scope of the present paper. It is also still today subject to ongoing research and thus is likely to develop in the coming years. Yet, already the body of knowledge hereby discussed indicates that we cannot say the ETA to be scientifically valid, as the premise that there will someday be an intelligent machine the cognitive capacities of which exceed those of humans has been shown to contain illogicalities. From the perspective of ethical argumentation, this compromises the quality of the argument and raises a question: is the speaker sincerely believing in what they propose, or is the argument driven by motives other than contribution to a better mutual understanding of the topic? This is what we next discuss in more detail.

3.2 Truthfulness: a competition-based approach to artificial intelligence

According to Habermas [33], communicative rationality requires the speaker to express their motives sincerely, i.e., *truthfulness*. A comprehensive evaluation of truthfulness would thus require knowing the subjective intention of the speaker, which in public discourse is often challenging to analyse. Hence, we can only evaluate the truthfulness of the ETA to the extent of public statements and what can be inferred based on the linguistic choices of the speakers in their argumentation. Also, it needs to be noted that in the scope of the present paper, we have not analysed all available public arguments made on behalf of existential threats. Rather, we bring up through examples certain tendencies in an ongoing discussion that are of significance when evaluating the sincerity of participants in the ETA discourse.

Firstly, it is worth noting that several of the ETA proponents hold a position where they would be among those who benefit the most from AGI development. For instance, the CEO of OpenAI, Sam Altman, has voiced concerns over the existential risks of AGI [59], all while leading a company the mission of which is to develop such technology. The choice of AI companies to rely on the ETA could indeed arise primarily from economic incentives instead of genuine concerns for the safety of people: the more powerful the technology appears to be, the more funding it is likely to receive to be safely brought to market [e.g., [60]]. As another example, while signing the FLI's Open letter calling for a pause in AI development, it seems that Elon Musk was simultaneously investing in his own company advancing the development of AGI [61].

This power of the ETA to advance economic interests is related to controlling of the discourse around AI. It is important to note that AI moguls are seeking control over *both* sides of the discourse in shaping the positive *as well as* negative narratives the public eventually receives. This can be seen as a strategy to ensure that no matter what angle the media portrays, it advances the hype around AI technologies in a way favourable to the AI companies, as their representatives get to choose the message. This can be seen as an effective strategy for marketing and PR: As Nelkin [60] noted already in the 1990s, media are strongly dependent on corporate communication when describing the features of new products. This creates favourable conditions for attaining control of discourse:

> "Relying on corporate sources of information about new products, the media have adopted a corporate rhetoric, promoting applications and accepting,

[9] See, e.g., a blog article of Emily Bender, Professor of Linguistics, about the term artificial intelligence on Medium: https://medium.com/@emilymenonbender/opening-remarks-on-ai-in-the-workplace-new-crisis-or-longstanding-challenge-eb81d1bee9f.

> unreflectively, the assumptions of aggressive industry seeking an expanded market." (p. 46)

Economic motivation, however, is often concealed in the discussion around AI. This implies that the ETA is often lacking in truthfulness. Yet, when looking at the lobbying efforts of the big AI companies, it becomes clear that there is an intention towards a regulatory capture to shape global AI policy based on economic interests. As was noted by Perrigo [62], despite the numerous pro-regulation statements of OpenAI's Sam Altman, the company has been actively lobbying to water down certain parts of the upcoming European AI regulation. Such actions indicate that calls for investments in safeguards for AGI by the company representatives do not necessarily reflect their intention to limit the way AI can be developed but rather their aim to market their product more efficiently to gain investor interest.

Let us take another example, this time from the realm of governmental AI policy. In his speech for the General Assembly of the United Nations in September 2023, the Deputy Prime Minister of the UK, Oliver Dowden, spoke about the existential threats brought forth by a likely occurrence of superintelligent AI.[10] Dowden proceeds to seek approval for the position of the UK by presenting the following defence:

> "For those that would say that these warnings are sensationalist, or belong in the realm of science-fiction, I simply point to the words of hundreds of AI developers, experts and academics, who have said - and I quote: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." I do not stand here claiming to be an expert on AI, but I do believe that policy-makers and Governments ignore this expert consensus at the peril of all of our citizens."

Doing so, Dowden presents the ETA as an "expert consensus", omitting a significant body of knowledge that points to another direction. As discussed in this paper, it is true that many experts hold such views as Dowden mentions. Yet, a closer inspection of who participates in the discourse on the threats of AI in the UK sheds light on the interests that have been given a voice—and thus control—in determining how we perceive the urgency of the ETA.

In September 2021, the UK Government published its National AI Strategy,[11] which is a ten-year plan "to make Britain a global AI superpower". The strategy comes with the aims of (i) holding UK's position as a leader in AI, (ii) supporting a transition to an AI-enabled economy, and (iii) ensuring proper AI governance to encourage innovation. In September 2023, it disbanded its independent advisory board on the ethics of AI without public announcement [63]. Instead, the British government established a Frontier AI Task Force led by a technology investor, Ian Hogarth, who is known to believe that AGI—or, as he calls it in an article he wrote for Financial Times, "God-like AI"—is possible, if not likely to emerge at any moment [64]. According to their first progress report,[12] the task force has thus far concentrated on hiring a research team on AI safety consisting of AI experts with experience from DeepMind, Microsoft, Redwood Research, The Center for AI Safety, the Centre for Human Compatible AI and RAND corporation, the latter four of which are organisations promoting the longtermist view of AGI risk, endorsed by proponents of the ETA. In preparing for British AI regulation, it seems that the impact of the big AI companies is eminent [e.g., [21]], and similar tendencies seem to arise elsewhere, such as in the US [65]. This clearly exhibits a worrying over-reliance on business-led initiatives [e.g., [66]].

It thus seems that the UK AI policy is emphasising viewpoints of ETA proponents in a way that echoes the message of companies focusing on AGI development. It directs the resources dedicated to ethical AI development towards issues such as alleged existential risks of AGI development at the expense of currently persisting problems in everyday use of AI systems. Focusing primarily on competition neglects the public good as the main concern for democratic government. On the other hand, however, this shift makes appeal to ethical considerations—namely protecting the public against existential level risks—which can be seen as its claim to communicative rationality.

The problematic nature of an excessive consideration of business interests in politics was pointed out even by the neoliberal economist Milton Friedman already in 1970 [67]. Friedman's point is that business leaders are presumably experts in running their companies. However, nothing about running companies make business leaders experts on public matters. The same goes for the various experts of AI who wish to take on social responsibility by concentrating on what AI businesses see as existential risks resulting from

[10] Deputy Prime Minister Oliver Dowden's speech to the UN General Assembly: 22 September 2023: https://www.gov.uk/government/speeches/deputy-prime-minister-oliver-dowdens-speech-to-the-un-general-assembly-22-september-2023.

[11] National AI Strategy, Office for Artificial Intelligence, UK: https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf, p. 4.

[12] Available here: https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report.

their technology. They may be assumed to know very little about public matters. Yet we now witness leaders in AI business and research take on social responsibility with the British government embracing this activity. Friedman rightly said that such involvement by business in social responsibility is intolerable on grounds of political principle.

Thus, in evaluating the truthfulness of the ETA in this context, we can ask: is the motivation of the UK government representatives to protect people from potential harms caused by AI development, or is it perhaps to implement its AI strategy and create favourable conditions for the UK to create and extract wealth? Considering the aforementioned UK government strategic aim of becoming an "AI superpower", and its heavy investments in attracting AI companies to the UK, it is prudent to question the truthfulness of the ETA as a genuine concern for the safety of people.

Another element that puts into question the truthfulness of the ETA is that of an alleged inevitability. Examining again the example of Deputy Prime Minister Dowden's speech, the UK government calls for a surrender in front of technology and competitive forces. According to Dowden,[13], "the starting gun has been fired on a globally competitive race in which individual companies as well as countries will strive to push the boundaries as far and fast as possible." The question is: the boundaries of what? What are we competing over?

The proposed solution for an allegedly inevitable AGI development is to compete in studying this yet growing but soon-to-be superintelligence in order to understand it before it has a chance to wipe us out. So, in essence, this narrative assumes we (humanity) are in competition against both technology itself as well as each other in developing such technology. If in the process there are negative consequences for the public, these are negligible in view of the predicament of becoming extinct. What we cannot do, however, is to impose regulations that would slow us down in the fight for our collective survival. This narrative conflates competition strategy with communicative rationality; competition is now seen as the communicatively rational response to the problems that face us.

The AI Now Institute has followed the evolving rhetoric of a supposed AI arms race over the course of five years. As their report states, the competition rhetoric "crystallize[s] the notion of AI systems (and the companies that produce them) not merely as commercial products but foremost as strategic national assets" that need preferential treatment in the form of "greater state support for a specific kind of large-scale AI innovation."[14] The question of what exactly are we competing over is a question that has no answer in terms of the final state or product. In this context, what is important is where the development and adoption of a given technology first occurs. As economist Carlota Perez [68] writes,

> "[Technological] revolution has generally irrupted in the core country of the previously prevailing paradigm, and spreads there first and propagates to the periphery [...]. Whichever the core, the installation period is very much marked by the polarisation between the front-running country or countries, where the new industries are being deployed, and those areas of the world that are left out and falling behind." (p. 63–64.)

If a country or countries manage to become the core of a new technological revolution, they stand to benefit in relation to other countries. Such competition is a form of colonisation as it creates inequalities between nations; inequalities which result in less bargaining power for some. The nations which lose out are then subject to catering for the needs of consumers in core countries. Such *digital colonialism* has been noted by, e.g., Coleman [69] (see also Section 3.3 below).

Hence, examples such as those of the British government show signs of widespread adoption of effective altruism and longtermist discourse in determining where the resources in ensuring safe AI should be directed (namely, to AGI research and development). As a result, many organisations doing AI policy and directing AI development appear to be permitted to do the opposite of what Dowden voices in his speech, namely that "Tech companies must not mark their own homework." The homework of tech companies is marked as if it were correct, that is, it is marked on their terms; what is marked is what they have chosen to be marked.

It is essential to recognise that some proponents of the ETA may genuinely believe that developing AGI is possible and likely, and therefore a reason for concern regarding human extinction (which was most probably the case of, for example, the Google engineer who suggested LaMDA was already sentient [70]). Whether as a result of lobbying on part of AI companies or by persuasion of AI experts, ignorance of the lack of scientific validity regarding the possibility of AGI means there are likely people who are truthfully contributing to the ETA discourse. Without accessing the inner motivations of these speakers, only limited conclusions can be drawn based on the observations discussed above. Therefore, when evaluating the overall

[13] Deputy Prime Minister Oliver Dowden's speech to the UN General Assembly: 22 September 2023: https://www.gov.uk/government/speeches/deputy-prime-minister-oliver-dowdens-speech-to-the-un-general-assembly-22-september-2023.

[14] "US-China AI Race: AI Policy as Industrial Policy." A report by AI Now Institute, published on 11 April 2023: https://ainowinstitute.org/publication/us-china-ai-race.

truthfulness of the ETA, attention and critical evaluation is needed when evaluating the motivations of the speakers, as there is strong indication that the validity claim of truthfulness is often lacking.

3.3 Rightness: ETA in the Western normative context

We have arrived at the last of the validity claims: *rightness*. Rightness refers to the normative validity of an argument. The purpose of considering the normative validity of an argument is to assess whether a proposed action is "right in the given normative context" [33, p.306]. Considering normative validity is thus a way of providing legitimacy to any proposed argument and consequent actions [33, p. 307]. In the ETA discourse, this translates into a consideration of the normative legitimacy of the calls for regulation grounded in Western normative thought, and the longtermist ideology.

Firstly, the advocates of the ETA often call for multistakeholder governance and governmental regulation of AI systems (such as in the case of the Open Letter of FLI[15]). The normative premises of the argument often seem to arise from existing normative frameworks, such as different types of regulation, human rights and positive law. These references of the ETA discourse seem to rely primarily on Western normative traditions, which resonates with findings of prior research indicating that most guidelines and strategies for responsible use of AI have been emerging in the Global North, on Western values [71].

The emphasis on Western normative frameworks that are prevalent in the Global North can be perceived as problematic. For example, many existing Large Language Models (LLMs) that have accelerated the discussion around the potential for AGI are built in a way that they can be accessed and used globally. In part, it has been suggested to lead to empowerment of the local communities, as these products are brought available to everyone through mobile interfaces. Whether the benefits and the costs of building and using the models are distributed equally has, however, raised questions. For example, in order for the ex-non-profit OpenAI[16] to build its ChatGPT, the company has relied on exploitation of low-paid workers in Kenya to classify data [62]. Having big, Western AI companies taking over the digital infrastructure in the Global South has obvious and striking similarities with both past colonisation practices as well as digital colonialism discussed by, e.g. Kwat [72]. Similarly, MIT Technology Review has dedicated an entire series for discussing AI Colonialism,[17] such as AI-enabled mass surveillance in South Africa contributing to apartheid [73]. Moreover, as Abeba Birhane points out, AI systems developed in the Global North and exported elsewhere are not neutral but impose Western normative frameworks to other contexts—they are "sold as a way of helping people in underdeveloped nations, but it's often imposed on them without consultation, pushing them further into the margins" [74]. Indeed, as Adams [71] has noted, it seems that the current AI development has not led to decolonisation of AI technologies. Rather, it has followed the logic of colonial economies benefiting from the resources in the Global South, all while bringing wealth to their owners in the Global North.

From a discourse perspective, considering that the discourse around AI is constructive of the technology, lopsided concentration on Western normative frameworks risks leading to AI systems that are not legitimate in the Global South but still force it to adapt. The phenomenon is known in data economy literature as *data colonialism*, which has normalised the exploitation of humans through personal data [75]. Data colonialism is of ever higher relevance in the context of AI development, which comes with hugely unbalanced value exchange both between AI companies and the public, as well as the Global North and South. Such imbalance often leads to reinforced marginalisation [76].

When we remind ourselves of what Western societies consider as being right, the ETA discourse runs into further problems that also concern the legitimacy of development of technology that is being imposed by the West on its own people, as well as other nations. The reason for this is that Western societies operate under the normative ideology of liberalism enshrined in the Western legal tradition. Furthermore, Western nations have made it a priority to try and impose the liberalist framework globally through supranational institutes and trade agreements, such as the International Monetary Fund (IMF), and General Agreement on Tariffs and Trade (GATT)/World Trade Organization (WTO). Central to liberalism is the tension between the good of the many and the good of the individual. It is precisely such tensions that are witnessed in the ETA discourse when it is stated that the development of technology is inevitable (see 3.2). This tension can be traced back to the liberalist idea of the freedom of the individual, with the word "libertarian" occurring in the late eighteenth century as a term referring to the freedom of will as opposed to "necessarianism" (what would now be called determinism)

[15] "Pause Giant AI Experiments: An Open Letter": https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[16] OpenAI shifted in 2019 from non-profit to a 'capped profit' company with returns for investors currently capped at 100 times their investment: https://openai.com/blog/openai-lp.

[17] The topic of AI Colonialism on the MIT Technology Review's website: https://www.technologyreview.com/supertopic/ai-colonialism-supertopic/.

[77, p.9]. From this metaphysical discussion, an ideological and political use of the idea of the libertarian idea of free will spread quickly and has since been used to justify actions in the name of individual freedom and the need to protect it.

Recently, Christoph Menke—in line with Spinoza [78, p. 685]—has called to question the legitimacy of the legal justification of the right of the individual to autonomy and choice, arguing that law's acceptance of the natural right of humans cannot be normatively justified [79]. This is because, argues Menke, in presupposing the non-legal (natural rights) as its matter, law cannot distinguish its form from its matter. Law becomes 'the right of rights': 'The normativity of the modern right [Recht] of rights has the structure of a legalization of the natural [79, p.19]. Here every individual is to have *equal right* to strive towards their own personal ends to the best of their abilities (short of physical violence). In defending equal right for personal striving, modern law seems to have institutionalised the law of the jungle.

Habermas is aware of this lack of justification for the normative validity of individual freedom, argues Menke, and goes on to posit "a functional and systemic imperative" of considering the freedom of the individual: an imperative that has legitimacy because it has been generated by discursive means [79]. Nevertheless, when we consider rightness not as a functional imperative but as the normative validity of an argument, we see that the Western conception of the freedom of the individual is not justifiable by the criterion of rightness. As such, the question is: why would we in the West who do not belong to the 1%—let alone other parts of the world—agree to a functional premise of the law of the jungle; a premise not of equality, but of equal individual right? By what right is the Western conception of the freedom of the individual justified? And what right do (Western) countries have to impose their own functional premise of the freedom of the individual over and against their own public, let alone other countries?

Secondly, to evaluate the rightness of the ETA, we need to revisit the normative assumptions of longtermism, which is rooted in utilitarian foundations of maximising utility for the maximum number of people. Therefore, any non-utilitarian consequentialist would have a hard time swallowing the normative premise of longtermism, and thus that of the ETA. Moreover, the ability of longtermism to lead to long-term accumulation of utility in the first place has been contested. For example, Hyde [80] unravels several of the persisting issues in the longtermist argument, such as the difficulty (if not impossibility) to estimate the impacts of actions on the future to a sufficient extent, as well as the problem of ignoring current unhappiness in search for future happiness, leading to a "shorterm view of the longterm future" (p. 148). Tarsney shows that—even when looked at from the most "hospitable" of perspectives—the potential validity of longtermism is essentially tied to an assumption that we "accept expectational utilitarianism, and therefore do not mind premising our choices on minuscule probabilities of astronomical payoffs" [29, p. 195.] Consequently, directing a considerable proportion of our available resources today to the safe development of AGI suggested by the ETA advocates[18] can reasonably be questioned in terms of its normative validity.

Moreover, even if we ignored all the inconsistencies and accepted the premise of longtermism, we would still run into issues when looking at who will benefit from this future-oriented AGI development. As discussed above, the benefits of the current AI development do not seem to spread equally to everyone but, rather, systematically discriminate against people in vulnerable positions and minorities [1] while also favouring people in the Global North [71]. Therefore, the people who are suffering now and their descendants are in risk of having lesser chances to enjoy the fruits of the alleged AGI in the future. Therefore, ignoring the current inequalities and ethical issues, the longtermist ETA seems to disregard the normative value of justice and fairness, called for by several modern normative frameworks.

Overall, it seems that the ETA is tied to the Western tradition of the freedom of the individual and equal rights, as well as the longtermist ideology rooted in utilitarianism. Considering its global impacts and the role of workers in the Global South that contribute to the development, use and training of these technologies, it would be ethically justified to broaden this perspective and consider the wider normative contexts of where the stakeholders of the systems reside and to critically discuss the normative legitimacy of AI regulation based on individual freedom. Otherwise, AI development risks aggravating ethical issues arising from colonialist practices and marginalisation of certain groups of people (such as indigenous people and people of colour), as well as oppressive and colonialist practices imposed by AI companies on people at large. Furthermore, as the prevalent Western liberalism driving AI development is ethically questionable, the ETA discourse would benefit from going beyond Habermasian examination of the fit between an argument and its normative context to engage in deliberation over whether the normative system we have in place is indeed ethically justified. Similarly, as the normative basis of longtermism has been frequently questioned, there

[18] Such statements include, for example, the recent suggestion of the OpenAI CEO for the need of 5 to 7 trillion USD for chips needed to continue their pursuit in AGI (see, e.g., VentureBeat's piece of news on the matter: https://venturebeat.com/ai/sam-altman-wants-up-to-7-trillion-for-ai-chips-the-natural-resources-required-would-be-mind-boggling/), which led to an intense discussion of the extent of the demands (see, e.g., Gary Marcus's blog on Substack: https://garymarcus.substack.com/p/seven-reasons-why-the-world-should.

seems to be a need for further ethical justification of the ETA and its normative premise. Therefore, we argue that the ETA can be considered right, i.e., normatively legitimate only from the position of the liberalist account that assumes equal individual right as the basis for AI regulation, combined with the acceptance of expectational utilitarianism. As discussed above, this can and thus should be subjected to further critical examination and deliberation in order to find an argument that could not be rejected as lacking in rightness by the majority of people concerned with global AI technologies.

4 Discussion: societal and ethical implications of ETA in AI hype

We have now discussed the ETA through the lens of Habermas's validity claims of truth, truthfulness, and rightness. This analysis unveils several persistent and potential ethical and societal consequences.

First, the lack of scientific validity of the ETA itself contributes to the lack of quality in the information based on which people make decisions regarding AI systems. Misguiding existential presupposition can lead to poor decisions on the part of policymakers when it comes to regulation and other safeguards, potentially leading to ignoring use cases that could actually cause existential risks: which systems should we automate, and which not? How do we make sure we do not overly trust the existing AI systems (automation bias; see, e,g. [81]) as safety components of critical systems? It is safe to say that without scientific validity, the ETA discourse concerning the idea of intelligent machines is not something that should be pursued. The existential risk is not one that is posed by machines, but by humans. Just as we may use the atomic bomb to wipe ourselves off the face of the earth, we can put AI in charge of deploying the bomb, and in that case AI would be an existential risk. AI is not capable of posing an existential risk by itself any more than any other material object or resource around us.

Second, in primarily focusing on long-term risks, the ETA neglects contemporary ethical and societal issues arising from existing AI systems, which could lead to harmful disruption of societal structures and people's lives. For instance, current technologies enabling widespread and effective mis- and disinformation are challenging democratic governance, which should be based on high quality public deliberation, freedom of opinion formation, and free, competitive elections [82, 83]. A recent incident in Slovakia, where fact-checkers performed poorly when trying to recognise fake audio recordings only days before a parliamentary election [84], serves as an example of the disruptive impact of the use of generative AI in determining power relations in society. On an individual level, many of the current issues in AI systems, such as algorithmic bias, concern minorities and people in vulnerable positions [85, 86]. Concentrating on long-term issues might lead to an existential risk to minorities much earlier, arising not from the technology but from how we as humans act on it, when AI systems poorly adapted to minorities are used to make decisions that have a profound impact on people's lives. Although these risks realise themselves in the near term, their impacts extend to the long term if not properly mitigated.

The narrative of inevitability attached to AGI development and the following alleged existential threat is equally problematic. Accordingly, technology is portrayed as an external artefact that is imposed on us regardless of our actions. This proposition further undermines the scientific validity, as well as raises questions on the truthfulness, of the ETA argument. Systems with components of AI are built by humans, for the use of humans. Our own politics, policies, economics, and law are firmly in control of technology. It is thus humans, not external forces (and certainly not machines) who hold the agency and allow the advancement of technology and build legal and social structures to enable it. In endorsing the ETA discourse while adopting effective altruism and longtermism, the speaker is attempting to portray competition as a moral duty. In so doing, the hope seems to be that one could lay claim to ethical conduct without engaging in any ethically good activity. This is a discourse around technology the aim of which is to construct a technology for economic purposes, not for the good of humanity.

From an ethics perspective, untruthfully driving one's economic interest over striving towards good life for all implies reliance on a strategic game driving economic interests instead of communicative action. From a Habermasian perspective, AI development that is mainly based on a strategic game cannot be ethical, as communicative action is a prerequisite for ethical action. As the ETA can be rejected on the basis of all three validity claims, a question remains on the intentions of the proponents of the ETA and the following ethicality of their actions. For example, the now-again-CEO of OpenAI, Sam Altman, did not seem hesitant to continue developing AGI for a profit-driven tech giant Microsoft, when he was temporarily fired from the capped-profit AI company.[19]

Lastly, it seems that there is a need for further inspection of the very normative validity of the ETA, as it is mainly

[19] Sam Altman was ousted by the OpenAI board for a period of five days, during which he was hired by Microsoft to continue developing AGI. See, e.g., a piece of news on Le Monde: https://www.lemonde.fr/en/pixels/article/2023/11/20/microsoft-hires-former-openai-chief-sam-altman_6271437_13.html.

based on a Western normative framework. As the AI systems developed by Western tech giants benefit from resources of the Global South and are used globally, imposing a Western normative framework should be questioned. Moreover, the emphasis on positive law based on individual freedom and equal rights of the individual requires further critical evaluation: even though legal, is the Western legal tradition leading to ethical action? This is a question that would merit further research also in the context of AI development.

For decades, there has been a tendency for the media to balance opposing positions when reporting about technology and science—one overwhelmingly optimistic and one pessimistic, if not apocalyptic—which has come with a cost of neglecting the root causes and technological explanations of these opposing views [60, p.48]. The aim of doing so has been to balance perspectives and interests of different societal actors. In today's discussion on AI, the extreme ends of the discussion gaining attention—existential threats and the tremendous opportunities—seem to both originate from the same source: those benefiting most in monetary terms. The challenges discussed above regarding the ETA discourse show that we need to direct more attention towards the impacts and the quality of the discourse around AI systems. In the light of Habermas's theory, it does not seem possible that the current AI hype could contribute to ethical development of AI, as long as arguments such as the ETA fail at least partially not just in one but in all validity claims—truth, truthfulness, and rightness.

5 Conclusions

In this paper, we have used Jürgen Habermas's critical theory of communicative action to analyse the current hype around AI technologies, focusing on the Existential Threat Argument (ETA). The framework furnishes us with three criteria with which to assess the discourse, namely truth, truthfulness, and rightness. We have seen that the ETA fails in each one of these, which indicates weak scientific and normative validity, as well as untruthfulness of the proponents of the argument. What cannot easily be known is which actors are engaging in the ETA discourse in ignorance and which actors are engaged in strategic games. Even in the former case, the ignorance can be countered: we do not have to concede the idea that no one knows the possibilities of AI, or that AI might become intelligent.

First, there is lack of truth by way of a lack of scientific validity concerning the intelligence of AI. Second, the unquestioned normative premise of liberalism and the emphasis on Western normative frameworks also renders its normative validity questionable. This means that the proponents of the ETA discourse lack rightness insofar as the normative context in which they operate is that of free individuals striving for their own interests in free markets based on the exchange of private property. Finally, the ETA discourse also lacks truthfulness. This is clear when the narrative portrays technological advancement as inevitable and calls for a surrender in front of competitive forces. In doing so, the ETA discourse portrays technological solutionism and competition as communicative rationality, that is, as something that has a basis in truth and rightness instead of strategic games. This is further emphasised in the way in which the AI companies contribute to both hype of the positives and the negatives of the technology, gaining control over both sides of the discourse.

In conclusion, the ETA lacks the demands of communicative rationality, which is likely to negatively impact the ethicality of AI development. This is apt for accommodating the strategic games of various stakeholders, according to their own interests. However, if we want to endorse ethical AI, there is an urgent need for redirecting communication around AI from the hype of ETA towards rational and transparent communication. Else, we will be reduced to rhetorical sophistry instead of concentrating on the substance of arguments.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., Kuk, G.: Will algorithms blind people? the effect of explainable ai and decision-makers' experience on ai-supported decision-making in government. Soc. Sci. Comput. Rev. **40**(2), 478–493 (2022). https://doi.org/10.1177/0894439320980118
2. König, P.D., Wenzelburger, G.: Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. Gov. Inf. Q. **37**(3), 101489 (2020). https://doi.org/10.1016/j.giq.2020.101489
3. Kazim, E., Koshiyama, A.S., Hilliard, A., Polle, R.: Systematizing audit in algorithmic recruitment. J. Intelligence **9**(3), 46 (2021)

4. Tilmes, N.: Disability, fairness, and algorithmic bias in ai recruitment. Ethics Inf. Technol. **24**(2), 21 (2022)
5. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
6. Rillig, M.C., Ågerstrand, M., Bi, M., Gould, K.A., Sauerland, U.: Risks and benefits of large language models for the environment. Environmental Science & Technology **57**(9), 3464–3466 (2023). https://doi.org/10.1021/acs.est.3c01106
7. Zuboff, S.: Big other: surveillance capitalism and the prospects of an information civilization. J. Inf. Technol. **30**(1), 75–89 (2015). https://doi.org/10.1057/jit.2015.5
8. Zuboff, S.: The Age of Surveillance: the Fight for a Human Future at the New Frontier of Power. Public Affairs, New York (2019)
9. Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A.: The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat. Med. **24**(11), 1716–1720 (2018). https://doi.org/10.1038/s41591-018-0213-5
10. Mirbabaie, M., Stieglitz, S., Frick, N.R.: Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Heal. Technol. **11**(4), 693–731 (2021). https://doi.org/10.1007/s12553-021-00555-5
11. Savaget, P., Chiarini, T., Evans, S.: Empowering political participation through artificial intelligence. Science and Public Policy **46**(3), 369–380 (2019). https://doi.org/10.1093/scipol/scy064
12. Bostrom, N.: The control problem. excerpts from superintelligence: Paths, dangers, strategies. Science Fiction and Philosophy: From Time Travel to Superintelligence, 308–330 (2016) https://doi.org/10.1002/9781118922590.ch23
13. Schopmans, H.R.: From coded bias to existential threat: Expert frames and the epistemic politics of ai governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 627–640 (2022). https://doi.org/10.1145/3514094.3534161
14. Swanson, E.B., Ramiller, N.C.: The organizing vision in information systems innovation. Organ. Sci. **8**(5), 458–474 (1997). https://doi.org/10.1287/orsc.8.5.458
15. Van Dijk, T.A.: Discourse and Power. Palgrave Macmillan, New York (2008)
16. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023)
17. Bostrom, N.: Superintelligence: Paths, Dangers. Strategies. Oxford University Press, United Kingdom (2014)
18. Tegmark, M.: Lif3.0: Being Human in the Age of Artificial Intelligence. Penguin, Great Britain (2017)
19. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin, United States of America (2019)
20. Manacourt, V., Scott, M., Goujard, C., Bordelon, B.: How rishi sunak convinced the world to worry about ai. Politico (2023). https://www.politico.eu/article/rishi-sunak-convince-world-worry-artificial-intelligence-ai/
21. Hern, A., Kiran, S.: No 10 acknowledges 'existential' risk of ai for first time. The Guardian (2023). https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak
22. Clarke, L.: How silicon valley doomers are shaping rishi sunak's ai plans. Politico (2023). https://www.politico.eu/article/rishi-sunak-artificial-intelligence-pivot-safety-summit-united-kingdom-silicon-valley-effective-altruism/
23. Guest, P.: Britain's big ai summit is a doom-obsessed mess. Wired (23.10.2023). https://www.wired.co.uk/article/britains-ai-summit-doom-obsessed-mess
24. Gabriel, I.: Effective altruism and its critics. J. Appl. Philos. **34**(4), 457–473 (2017). https://doi.org/10.1111/japp.12176
25. Singer, P., MacAskill, W.: Introduction, p. Center for Effective Altruism, Oxford (2015)
26. Ioannidis, I.: Shackling the poor, or effective altruism: A critique of the philosophical foundation of effective altruism. Conatus 5(2), 25 (2020) https://doi.org/10.12681/cjp.22296
27. Greaves, H., Pummer, T.: Effective Altruism: Philosophical Issues. Oxford University Press, Oxford (2019)
28. Greaves, H., MacAskill, W.: The case for strong longtermism. University of Oxford, Global Priorities Institute (2021)
29. Tarsney, C.: The epistemic challenge to longtermism. Synthese **201**(6), 195 (2023). https://doi.org/10.1007/s11229-023-04153-y
30. Caterino, B., Hansen, P.: Critical Theory, Democracy, and the Challenge of Neo-Liberalism. University of Toronto Press, Toronto (2019)
31. Delanty, G., Harris, N.: Critical theory and the question of technology: The frankfurt school revisited. Thesis Eleven **166**(1), 88–108 (2021). https://doi.org/10.1177/07255136211002055
32. Habermas, J.: Theory and Practice. Beacon Press, Boston, Mass (1973)
33. Habermas, J.: The Theory of Communicative Action, vol. 1. Polity Press, Cambridge (1984)
34. Habermas, J.: Justification and Application. Polity Press, ??? (1993). Translated by Cronin, Ciaran
35. Habermas, J.: Between facts and norms, trans. william rehg. Polity, 274–328 (1996)
36. James, M.R.: Communicative action, strategic action, and intergroup dialogue. Eur. J. Polit. Theo. **2**(2), 157–182 (2003). https://doi.org/10.1177/1474885103220003
37. Robinson, W.: Epiphenomenalism. The Stanford Encyclopedia of Philosophy (Summer 2023 Edition) (2023). https://plato.stanford.edu/archives/sum2023/entries/epiphenomenalism/
38. Chalmers, D.J.: The Conscious Mind. In Search of a Fundamental Theory. Oxford Paperbacks, Oxford (1997)
39. Metzinger, T.: The ego Tunnel: The Science of the Mind and the Myth of the Self. Basic Books (AZ), New York (2009)
40. Damasio, A.R.: The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Houghton Mifflin Harcourt, London (1999)
41. Crick, F.: The Astonishing Hypothesis. Touchstone London, London (1995)
42. Edelman, G.M.: Bright Air. Brilliant Fire. BasicBooks, New York (1992)
43. Baars, B.J.: In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press, Oxford (1997)
44. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., : Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708 (2023) https://doi.org/10.48550/arXiv.2308.08708
45. Minsky, M.: Society of Mind. Simon and Schuster, New York (1988)
46. Russell, S.J., Norvig, P.: Artificial Intelligence a Modern Approach, 4th edn. Pierson, Great Britain (2022)
47. Aru, J., Larkum, M.E., Shine, J.M.: The feasibility of artificial consciousness through the lens of neuroscience. Trends Neurosci. (2023). https://doi.org/10.1016/j.tins.2023.09.009
48. Bennett, M.R., Hacker, P.M.S.: Philosophical Foundations of Neuroscience. Blackwell Publishing, Oxford (2003)

49. Mingers, J.: Embodying information systems: the contribution of phenomenology. Inf. Organ. **11**(2), 103–128 (2001). https://doi.org/10.1016/S1471-7727(00)00005-1
50. Husserl, E.: Ideas: General Introduction to Pure Phenomenology. Routledge, London (2012). Translated by Moran, Dermot (2012), originally published in 1913
51. Heidegger, M.: Sein und Zeit, (1927). Translation Oleminen ja Aika by Kupiainen R. 2000. Tampere: Vastapaino
52. Merleau-Ponty, M.: Phenomenology of Perception. Routledge, London (1945). Translated by Donald A. Landes. 2012 London and New York: Routledge
53. Habermas, J.: The Theory of Communicative Action: Lifeworld and Systems, a Critique of Functionalist Reason, vol. 2. Polity Press, Cambridge (1987)
54. Fairtlough, G.H.: Habermas' concept of "lifeworld". Systems practice 4, 547–563 (1991) https://doi.org/10.1007/BF01063113
55. Montemayor, C.: Language and intelligence. Mind. Mach. **31**(4), 471–486 (2021)
56. Browning, J., LeCun, Y.: Language, common sense, and the winograd schema challenge. Artificial Intelligence, 104031 (2023)
57. Browning, J.: Personhood and ai: Why large language models don't understand us. AI & SOCIETY, 1–8 (2023)
58. Watson, D.: The rhetoric and reality of anthropomorphism in artificial intelligence. Mind. Mach. **29**(3), 417–440 (2019). https://doi.org/10.1007/s11023-019-09506-6
59. Kelly, S.: Sam altman warns ai could kill us all. but he still wants the world to use it. CNN Business (2023). https://edition.cnn.com/2023/10/31/tech/sam-altman-ai-risk-taker/index.html
60. Nelkin, D.: Selling Science: How Press Covers Science and Technology. W. H. Freeman and Company, New York (1995)
61. Heikkila, M.: What's changed since the "pause ai" letter six months ago? MIT Technology Review (2023). https://www.technologyreview.com/2023/09/26/1080299/six-months-on-from-the-pause-letter/
62. Perrigo, B.: Exclusive: Openai used kenyan workers on less than \$2 per hour to make chatgpt less toxic. Time Magazine (2023). https://time.com/6247678/openai-chatgpt-kenya-workers/
63. Martin, A.: British government quietly sacks entire board of independent ai advisers. The Record (2023). https://therecord.media/uk-disbands-ai-advisory-board-cdei-rishi-sunak
64. Hogarth, I.: We must slow down the race to god-like ai. Financial Times (2023). https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2
65. Bordelon, B.: Think tank tied to tech billionaires played key role in biden's ai order. Politico (2023). https://www.politico.com/news/2023/12/15/billionaire-backed-think-tank-played-key-role-in-bidens-ai-order-00132128
66. Rajvanshi, A.: Rishi sunak wants the u.k. to be a key player in global ai regulation. Times (2023). https://time.com/6287253/uk-rishi-sunak-ai-regulation/
67. Friedman, M.: The social responsibility of business is to increase its profits. In: Corporate Ethics and Corporate Governance, pp. 173–178. Springer, Germany (2007). https://doi.org/10.1007/978-3-540-70818-6_14
68. Perez, C.: Technological Revolutions and Financial Capital. Edward Elgar Publishing, Cheltenham (2003)
69. Coleman, D.: Digital colonialism: The 21st century scramble for africa through the extraction and control of user data and the limitations of data protection laws. Mich. J. Race & L. 24, 417 (2018). https://heinonline.org/HOL/P?h=hein.journals/mjrl24 &i=429
70. Tiku, N.: The google engineer who thinks their ai is alive. The Washington Post (2023). https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/
71. Adams, R.: Can artificial intelligence be decolonized? Interdisciplinary Science Reviews 46(1-2), 176–197 (2021) https://doi.org/10.1080/03080188.2020.1840225
72. Kwet, M.: Digital colonialism is threatening the global south. Al Jazeera (2019). https://www.aljazeera.com/opinions/2019/3/13/digital-colonialism-is-threatening-the-global-south
73. Hao, K., Swart, H.: South africa's private surveillance machine is fueling a digital apartheid. MIT Technology Review (2022). https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid/
74. Browne, G.: Ai is steeped in big tech's 'digital colonialism'. Wired (2023). https://www.wired.co.uk/article/abeba-birhane-ai-datasets
75. Couldry, N., Mejias, U.A.: Data colonialism: Rethinking big data's relation to the contemporary subject. Television & New Media **20**(4), 336–349 (2019). https://doi.org/10.1177/1527476418796632
76. Arora, A., Barrett, M., Lee, E., Oborn, E., Prince, K.: Risk and the future of ai: algorithmic bias, data colonialism, and marginalization. Inf. Org. **33**(3), 100478 (2023). https://doi.org/10.1016/j.infoandorg.2023.100478
77. Zwolinski, M., Tomasi, J.: The Individualists: Radicals, Reactionaries, and the Struggle for the Soul of Libertarianism. Princeton University Press, New Jersey (2023)
78. Spinoza, B.D.: Spinoza: The Complete Works. Hackett Publishing, Indianapolis (2002)
79. Menke, C., Turner, C.: Critique of Rights. Polity, Cambridge (2020)
80. Hyde, B.V.E.: The problem with longtermism. Ethics in Progress **14**(2), 130–152 (2023)
81. Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., Gray, K.: Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. Gen. **152**(1), 4 (2023). https://doi.org/10.1037/xge0001250
82. Kilovaty, I.: Legally cognizable manipulation. Berkeley Tech. LJ 34, 449 (2019). https://heinonline.org/HOL/P?h=hein.journals/berktech34 &i=491
83. Manheim, K., Kaplan, L.: Artificial intelligence: Risks to privacy and democracy. Yale JL & Tech. 21, 106 (2019). https://heinonline.org/HOL/P?h=hein.journals/yjolt21 &i=106
84. Maker, M.: Slovakia's election deepfakes show ai is a danger to democracy. Wired UK (2023). https://www.wired.co.uk/article/slovakia-election-deepfakes
85. Kordzadeh, N., Ghasemaghaei, M.: Algorithmic bias: review, synthesis, and future research directions. Eur. J. Inf. Syst. **31**(3), 388–409 (2022). https://doi.org/10.1080/0960085X.2021.1927212
86. Nishant, R., Schneckenberg, D., Ravishankar, M.: The formal rationality of artificial intelligence-based algorithms and the problem of bias. J. Inf. Technol. 02683962231176842 (2023). https://doi.org/10.1177/02683962231176842