ALGORITHM WATCH

/ Position Paper

Focus Attention on Accountability for AI – not on AGI and Longtermist Abstractions

Michele Loi, with Angela Müller and Matthias Spielkamp

September 2025

Current debates on AI are polarized by extreme optimism and pessimism.

Too many current debates about Artificial Intelligence (AI) are driven by extreme positions, both optimistic and pessimistic. Many tech CEOs and influential scientists hail AI as a savior for humanity, others as an existential threat. At AlgorithmWatch, we take current, short-term, mid-term, as well as long-term risks from the development and use of highly powerful AI systems seriously.

That said, we also see analytic weaknesses that undermine effective AI governance when extreme positions like "superintelligence" or "Artificial General Intelligence (AGI)" receive outsized attention in the debate, while highly capable models already demand urgent governance.

AlgorithmWatch takes long-term risks seriously but rejects AGI as the sole focus.

The work of AlgorithmWatch is focused on real, current, and immediate challenges and benefits from AI that manifest in the present or the near future.

We believe in also considering long-term risks to society from AI, and in ensuring we use evidence to achieve genuine positive impact. These two positions are also held, respectively, by schools of thought known as “longtermism” and “effective altruism.” However, in our view, these two schools of thought show gaps and weaknesses when applied to the social and political impacts of AI. They have also attracted powerful supporters and receive outsized resources, as they align well with Silicon Valley’s technological ambition, ideological underpinnings, and allied political activity, especially from the current US administration.

By highlighting these criticisms and weaknesses, we demonstrate why we at AlgorithmWatch take a different, and we argue more appropriate, approach to priority setting.

Philosophical weaknesses of AGI, EA, and Longtermist perspectives

The philosophical weaknesses are, in brief:

- **Technological determinism**: Narratives about AI systems developing into some form of "superintelligence" or "AGI" assume technology ever progresses inevitably. In this way, the speculative assumption about the future existence of AGI (with some degree of probability) becomes the given assumption. This (non-evidence-based) assumption then influences all other policy choices as a constraint, instead of being one policy outcome that may be promoted or avoided like others.
- **False precision and uncertainty**: Mimicking the rigor of mathematics in moral and political arguments is a mere pretense of theoretical rigor as it masks our ignorance and uncertainty about the future. However, the apparent clarity can be appealing to powerful decision-makers.
- **Illusion of neutrality**: Related to the above, the abstractions focus on long-term, low-disagreement, high-stakes risks (namely the continued existence of humanity). This can seem neutral, but actually directs attention onto imagined catastrophes with large hypothetical numbers, away from harder-to-quantify but immediate impacts. This also avoids serious confrontation with legitimate democratic questions and complex trade-offs which go beyond a technological scope.

Definitions

- **AGI/superintelligence**: Hypothetical systems with general cognitive abilities (for superintelligence: exceeding those of humans across most domains). These should not be treated as an inevitable future of AI.
- **Effective Altruism (EA)**: A contemporary approach advocating an evidence-based approach in matters of beneficence, coupled with the widespread belief by its practitioners that such an approach should be characterized by impartial calculations (as exemplified by utilitarianism) and a narrow view of human reason, derived from quantitative or formal frameworks in decision theory and economics.
- **Longtermism**: The idea that far-future effects of our actions (how they influence the probability of extinction, for example) should dominate decision-making relative to present concerns (the probability of human extinction is given more weight than present-day improvements, like building a hospital, for example).

Technological Determinism in AI Discourse

Many leading proponents of AI treat the development of superintelligence/AGI as inevitable. “We are now confident we know how to build AGI as we have traditionally understood it,” declared OpenAI CEO Sam Altman,[1] claiming that AGI could be achieved in 2025 and that it is now simply an engineering problem.[2]

As the highly influential longtermist philosopher William MacAskill and effective altruist Fin Moorhouse explain: “At the moment, the effective cognitive labor from AI models is increasing more than twenty times over, every year… For business to go on as usual, then current trends – in pre-training efficiency, post-training enhancements, scaling training runs and inference compute – must effectively grind to a halt.”[3]

Many such arguments also claim that, if we are destined to achieve AI with power equal to the human mind, this can be used to further improve AI and at that point, an era of unimaginable AI powers and risks inevitably emerges.

In the positive, optimistic framing put about by many proponents of unconstrained AI development, the coming of AGI is seen as the moment at which all questions of political governance are reconfigured; many new solutions become possible as AGI puts at the disposal of humanity large quantities of almost instantaneous intellectual work. As MacAskill and Moorehouse write in *Preparing for the Intelligence Explosion*, “one tempting plan is to ensure superintelligent AI is aligned, then rely on it to solve our other problems. This does make sense in many cases.” Although they do not mention the environment, others have used the example that AI will help us develop commercially viable fusion power, and all environmental challenges will sort themselves out.[4]

In the pessimist scenario in which AGI level or superintelligent agents are not aligned with human goals, radically different governance problems emerge. This scenario implies an unprecedented level of risk: a concrete possibility of human extinction or subjugation that dwarfs the significance of all other problems. Moreover, proponents say, this scenario needs to be prevented as an urgent priority rather than reacted to, as the combined speed and sophistication of superintelligent responses makes them unbeatable. As MacAskill and Moorehouse continue in their analysis, “some challenges will arise before superintelligence can help, and some solutions need to be started well in advance.”[5]

This is evident in how resources are allocated. Since 2023, OpenAI has launched grants specifically for “technical research towards the alignment and safety of superhuman AI systems,” while Open Philanthropy has directed $336 million to AI safety research focused on alignment.[6] The UK's Alignment Project, backed by “an international coalition of governments, industry, venture capital and philanthropic funders,” including the UK Security Institute and the Canadian AI Safety Institute, aims to ensure “AI

systems operate according to our goals" because "today's methods for training and controlling AI systems are likely insufficient for the systems of tomorrow."[7]

In this way, the *AGI and/or superintelligence as destiny* framing becomes the most urgent question for humanity, above those of democratic governance, planetary boundaries, and corporate accountability. In the above analysis, for example, long-term problems like environmental degradation due to intense energy use in the race to AGI (and then superintelligence) are not mentioned once. But what if the purported AI-driven development of fusion does not come fast enough, or is insufficient? Or what if the AI solutions help address overall energy use, but not differential impacts like water usage?[8] What agency does this framing give citizens and their governments? Instead of asking what kinds of AI systems we want (and by whom they should be developed under what constraints), the debate becomes about how to align superintelligence with human values, given that it is coming.

While near-term accountability work and long-term safety research could theoretically inform each other, this complementarity breaks down when longtermist frameworks treat existential risks as categorically more important than present harms. This follows directly from:

- expected value calculations that multiply across infinite future generations,
- cost/benefit analysis that does not allow for recognizing deontological principles (e.g., human rights) that are oriented at values such as dignity and autonomy,
- the treatment of uncertainty through formalized decision theory that obscures its speculative nature,
- the institutional dynamics where those controlling the "long-term" narrative also mobilize significant resources.

As long as these frameworks dominate, genuine dialogue remains impossible. The longtermist position cannot recognize present harms as comparably important without abandoning its effective-altruist and formalist core quantitative methodology. Until longtermist approaches abandon the risk assessment methodologies that assign overwhelming expected value to preventing low-probability extinction events, near-term and long-term perspectives remain structurally at odds – not because they are incompatible, but because one framework is designed to override the other.

False Precision and Uncertainty

Frameworks that have been developed to gauge uncertainty and estimate future outcomes are being used problematically, in a way that supports the priorities set by companies even without relying on the technologically determinist lens. This approach is most prominently seen in quantitative risk assessment frameworks that underpin effective altruism and longtermism, which combine probability estimates and expected

value calculations, leading to the thesis that the distant future consequences of our actions should dominate decision-making.[9]

The mathematical logic is simple but problematic: Multiply even a 1% chance of a consequence that influences virtually infinite future generations, and the value of that consequence is magnified in a way that no consequence only affecting the present can match. So, for example, preventing AI extinction becomes humanity's top priority regardless of other urgent needs. Once existential risk is on the table, no matter how speculative the probability that a current action may cause it, any other concern with mere present consequences, e.g., respect for human rights of current generations, can be dismissed as unimportant by comparison.

This approach trades on false certainty. One problem is that gauging uncertainty conflates two different types of probability. Expected utility theory ideally works with objective probabilities – frequencies observed through repeated trials, like coin flips. But AGI probabilities are subjective assessments collected by polling experts about unprecedented developments. When researchers survey AI scientists about the probability of human-level AI by 2030, they collect personal judgments, not objective data.[10]

The second aspect of false certainty concerns potential biases around causal pathways. Longtermist frameworks may systematically underweight risks with complex causal chains. The chain <human rights violations → democratic erosion → coordination failure → inability to address existential threats> represents a comparably plausible pathway to civilizational collapse as AI misalignment, yet receives far less longtermist attention. This creates a methodological blind spot where risks with shorter, more model-able pathways get prioritized over other dangerous institutional risks. This argument would require empirical research to further elaborate the prevalence and impacts of such biases in detail, but it is important to consider how frameworks claiming to optimize for long-term flourishing of the human race may systematically undervalue the institutional foundations that make flourishing of society and democracy possible.

Empirical-sounding estimations of AI risk now appear in the writings of philosophers, seemingly in the service of concrete agenda-setting goals.[11] Clearly, their scientific authoritativeness is highly disputable. But it has real consequences: By treating speculative future disasters as mathematically calculable, it redirects attention and resources away from AI's concrete harms today toward imagined catastrophes tomorrow.

The Illusion of Neutrality

Longtermism appears plausible because it focuses on outcomes that almost everyone agrees are bad, and effective altruism frameworks give this narrative a veneer of

neutrality. Human extinction, domination by unaligned AI, and loss of future potential seem like universal concerns. Combined with claims that future generations matter as much as present ones, this produces seemingly overwhelming reasons to focus on existential risk.

If these approaches get formalized into mathematical frameworks, they appear neutral but actually favor certain stances. The mathematics makes it seemingly easy to calculate massive future benefits (preventing extinction of trillions, almost an infinite value) but implausible to weigh in present moral concerns such as human dignity or justice as something comparable to extinction. Contrary to what eminent longtermist thinkers claim, Kantian “moral worth” does not scale into a calculable quantity that can swamp current duties.[12]

In other words, the illusion of neutrality consists in the fact that other moral views and intuitions are implicitly down-weighted – not because they are less valid, but because they are less mathematically tractable. Certain frameworks look like balanced synthesis but work as a device to unleash utilitarian dominance over other ways of thinking about ethics, such as more rights-oriented frameworks where violations of basic rights cannot simply be outweighed by benefits elsewhere.[13]

In concluding we should note that we can distinguish between longtermism's valid moral claim – that future individuals matter, and we should take steps to ensure a decent world for them – and the problematic intellectual apparatus that accompanies it. These features are not essential to caring about the future (longtermism as a concept in the broad sense); they rather emerge from a particular approach to moral mathematics.

Narrative Influence

The issues raised above can, unfortunately, be seen as positives to both ambitious Silicon Valley companies and many politicians and government officials. The longtermist and effective altruist approaches we have discussed above claim neutrality, broad agreement on key problems, and the possibility of technical solutions rather than the harder task of democratic legitimation and consequent trade-offs.

Longtermist ethics conveniently focuses on risks that don't threaten current business models or investor returns. Whether superintelligence emerges or not, companies face no near-term financial consequences – either AI remains controllable (no problem for profits) or becomes uncontrollable (profits become irrelevant amid human extinction). AI companies can present themselves as ethically responsible through safety research while avoiding accountability for more immediate harms to sustainable development. Even immediate risks, such as the threat that AIs may pose to the autonomy of human users through emotional manipulation, can be neglected when long-term existential risk research consumes a substantial portion of alignment funding.[14]

We have already listed some examples of how narrative influence appears in practice, but there are more. Influential organizations in the AI policy space include the Future of Life Institute and Open Philanthropy (which grew out of GiveWell Labs, founded by effective altruists).[15] There has been a general encroachment of effective altruist ideas, personnel, and money into political and philanthropic spaces, as tracked for example by *Politico* and the European AI and Society Fund.[16] Open Philanthropy's Technology Policy Fellowship placed Max Katz with Senator Martin Heinrich in 2022. Heinrich subsequently became one of only four senators selected by Senate Majority Leader Chuck Schumer to lead the US Senate's comprehensive AI policy development through year-long "Insight Forums" throughout 2023-2024. This demonstrates direct institutional influence channels between effective altruist organizations and key AI policymakers.[17]

The AlgorithmWatch Alternative

AlgorithmWatch offers an alternative grounded in democracy, rights, and lived impacts.

At AlgorithmWatch, we recognize there are both immediate and long-term risks of AI, including from increasingly powerful models. We are open to borrowing quantitative, systematic, and/or forecasting approaches from fields more broadly, beyond the constraints of narrow longtermism – as long as they also keep the focus on the issues we outline above and the moral principles which we adhere to (that are based on justice and autonomy), and don't let them get lost in numbers, abstraction, and narrow or elite-driven technical solutions.

We choose a rights-first, participatory, and evidence-led alternative that addresses concrete harms today and tomorrow while installing robust guardrails for the uncertain future. This means that we focus on imbalances of power between corporations, governments, and affected communities, on democratic legitimacy and inclusive governance processes, on localized impacts, shaped by contexts far from Silicon Valley, and on technologies that respect accountability and human agency rather than dictate governance by their own momentum.

This work includes, but is not limited to:

- developing tools for conducting Fundamental Rights Impact Assessments which are based on systematic methodologies to compare and evaluate risks, but which also explicitly address the limitations of such approaches and embed diverse input and evaluation into the methodology;
- conducting research into LLM uses which encompass both specific local contexts and broader risks – as seen, for example, in our work on safeguards against election misinformation in last years' regional elections in Germany and national

elections in Switzerland, where we highlighted how safeguards did not effectively recognize and protect regional political candidates;[18]

- journalistic and human-centered investigations which expose the real impacts of AI in particular contexts, whether through the expansion of data centers into many communities in Europe,[19] conditions of workers training AI,[20] or experiences of algorithmic discrimination[21] – to give just a few examples.

What is at stake is not whether humanity survives an abstract future intelligence, but whether we choose to govern the systems we are already building in ways that expand freedom, justice, autonomy, and sustainability rather than restrict it. The future of AI is not written in the stars but in parliaments, workplaces, and communities. Grand narratives about superintelligence may dazzle, but it is the quieter work of building accountability, sharing power, and protecting rights that will determine whether AI strengthens or undermines democracy.

This is where our attention belongs. If longtermist framings invite us to imagine eternity, our task is more immediate and no less ambitious: to ensure that today's decisions about AI leave behind institutions capable of facing whatever future arrives.

---

[1] Sam Altman, "Three Observations," 9 February 2025, personal blog. In particular "[t]he intelligence of an AI model roughly equals the log of the resources used to train and run it." https://blog.samaltman.com/three-observations. See also Sam Altman, "Reflections," 6 January 2025, personal blog, https://blog.samaltman.com/reflections. A useful summary of Altman's view is to be found in Tharin Pillay, "How OpenAI's Sam Altman Is Thinking About AGI and Superintelligence in 2025," *Time*, 8 January 2025, https://time.com/7205596/sam-altman-superintelligence-agi/.

[2] Often determinism takes on a more pessimist tone. See for example Eliezer Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut it All Down," *TIME*, 29 March 2023, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/. Peter Thiel presents a contrasting pessimistic perspective, arguing that technological development − including in AI − is actually stagnating across all domains: technological, economic, and social advocating for greater entrepreneurial freedom and warning against the "apocalyptic" threat of world government structures that could further constrain innovation. See: Ross Douthat, host, "A.I., Mars and Immortality: Are We Dreaming Big Enough?", Interesting Times podcast with Ross Douthat, episode featuring Peter Thiel, https://youtu.be/vV7YgnPUxcU.

[3] "Preparing for the Intelligence Explosion," MacAskill and Morehouse, https://www.forethought.org/research/preparing-for-the-intelligence-explosion.

[4] For example, Demis Hassabis (CEO, Google DeepMind), in an interview in 2025, challenged on the AI energy sustainability issue, said: "Yes, the energy required is going to be a lot for AI systems, but the amount we're going to get back, even just narrowly for climate [solutions] from these models, it's going to far outweigh the energy costs." *Guardian,* 4 August 2025, https://www.theguardian.com/technology/2025/aug/04/demis-hassabis-ai-future-10-times-bigger-than-industrial-revolution-and-10-times-faster. A similar statement appears in Sam Altman, on *X* (formerly Twitter), 7 May 2023, https://x.com/sama/status/1655249663262613507. These are examples of techno-optimist/accelerationist rhetoric that share longtermism's future-oriented ends with a more optimist spin – see, for example, Marc Andreessen, "The Techno-

Optimist Manifesto," *Andreessen Horowitz* (blog), 16 October 2023, https://a16z.com/the-techno-optimist-manifesto/.

[5] William MacAskill and Fin Morehouse, "Preparing for the Intelligence Explosion," 2025, https://www.forethought.org/research/preparing-for-the-intelligence-explosion. The piece sketches a broader framework of "grand challenges," including AI-enabled autocracies, destructive novel weapons, governance of digital beings, and off-world resource races – without addressing ecological risks such as environmental degradation from heavy energy usage or compute-intensive data center operations.

[6] "An Overview of the AI Safety Funding Situation," LessWrong, 2023, https://www.lesswrong.com/posts/WGpFFJo2uFe5ssgEb/an-overview-of-the-ai-safety-funding-situation

[7] UK Department for Science, Innovation and Technology (AI Security Institute), "AI Security Institute launches international coalition to safeguard AI development," government press release, 30 July 2025, https://www.gov.uk/government/news/ai-security-institute-launches-international-coalition-to-safeguard-ai-development; see also "The Alignment Project: About" page for partners and mandate, https://alignmentproject.aisi.gov.uk/.

[8] "With Google as My Neighbor, Will There Still Be Water?," AlgorithmWatch, 2023, https://algorithmwatch.org/en/protests-against-data-centers/

[9] "The Case for Strong Longtermism," Greaves & MacAskill, GPI Working Paper 2021, https://www.globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf.

[10] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," 2017, https://arxiv.org/abs/1705.08807. As critics have pointed out, "for existential risk from AI, there is no reference class" and "these estimates are not backed by any methodology." See Arvind Narayanan and Sayash Kapoor, "AI Existential Risk Probabilities Are Too Unreliable to Inform Policy," AI Snake Oil (Substack), 26 July 2024, https://www.aisnakeoil.com/p/ai-existential-risk-probabilities. Note that researchers attempting to quantify AI existential risk acknowledge their estimates as "highly-unstable" and "subjective," with one noting "a ~5% percent chance of existential catastrophe by 2070" while emphasizing the deep uncertainty involved. But this does not prevent them from influencing decision-making, despite the fact that the approach is ethically and scientifically controversial, as we discuss below. See "Draft Report on Existential Risk from Power-Seeking AI," Effective Altruism Forum, 28 April 2021, https://www.lesswrong.com/posts/HduCjmXTBD4xYTegv/draft-report-on-existential-risk-from-power-seeking-ai.

[11] For example, consider MacAskill and Moorehouse: "We can estimate a baseline by assuming that progress in training is entirely converted into improved inference efficiency, and more inference compute is entirely used to run more AIs… So if current trends continue to the point of human-AI parity in terms of research effort, then we can conclude AI research effort would continue to grow by at least 25x per year." Or: "The product of training compute, algorithmic efficiency, and inference compute—which combine to give AI research effort—will then have increased one hundred billion-fold (10^{11}), averaging just over 10x per year." In "Preparing for the Intelligence Explosion", 2025, https://www.forethought.org/research/preparing-for-the-intelligence-explosion.

[12] Philosophically, the expected choice-worthiness theory developed by MacAskill and Ord, two of the lead philosophical voices behind existential risk prioritization, falls prey to this criticism as well, but there is no space here for an extended argument on this matter. See MacAskill and Ord, 2020. "Why Maximize Expected Choice-Worthiness?" 1. Noûs 54, 327–353. https://doi.org/10.1111/nous.12264. Similarly their attempts to vindicate long-termism without assuming consequentialism in "The Case for Strong Longtermism" does not capture the way

deontologists think about ethics, as it fundamentally treats people or actions as bearers of value, not as ends in themselves. See "The Case for Strong Longtermism," GPI Working Paper June 2021, https://www.globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf.

[13] At least if rights function as side-constraints. See Robert Nozick, 1974. *Anarchy, State, and Utopia*, Basic Books, New York.

[14] Based on public job postings as of 29 Aug 2025, Anthropic and OpenAI show distinct hiring emphasis patterns. See Anthropic, "Frontier Red Team / RSP Evaluations," Anthropic Job Board, listing captured 29 August 2025, https://job-boards.greenhouse.io/anthropic.
OpenAI, "Safety Systems – Misalignment Research," *OpenAI Careers*, Greenhouse.io, listing captured 29 August 2025, https://openai.com/careers/senior-researcher-safety-systems-misalignment-research/. OpenAI, "Research Engineer / Research Scientist – Model Behavior," *OpenAI Careers*, job listing, published circa late August 2025, https://openai.com/careers/research-engineer-research-scientist-model-behavior/. Both companies appropriately prioritize CBRN (chemical, biological, radiological, nuclear) evaluations, however, beyond these justified safety measures, a qualitative inspection of current hiring patterns seem to reveal substantial resources directed toward more speculative scenarios. By contrast, both companies show fewer research-focused openings for "Societal Impacts" work, such as the Environmental & Supply Chain Responsibility Manager role at OpenAI. OpenAI, "Environmental & Supply Chain Responsibility Manager," OpenAI Careers, job listing, posted August 2025, https://openai.com/careers/environmental-management-supply-chain/.

[15] "An Overview of the AI Safety Funding Situation," LessWrong https://www.lesswrong.com/posts/WGpFFJo2uFe5ssgEb/an-overview-of-the-ai-safety-funding-situation.

[16] "Who's funding AI & Society work in Europe? A landscape review," European AI & Society Fund, https://europeanaifund.org/wp-content/uploads/2023/06/260623-FOR-PUBLICATION-EAISF-Funding-landscape-review.pdf , "When Silicon Valley's AI warriors came to Washington," *Politico* 30 December 2023, https://www.politico.com/news/2023/12/30/ai-debate-culture-clash-dc-silicon-valley-00133323, John Naughton "Longtermism: how good intentions and the rich created a dangerous creed," *Guardian* 4 December 2022 https://www.theguardian.com/technology/commentisfree/2022/dec/04/longtermism-rich-effective-altruism-tech-dangerous.

[17] For European context: This represents a form of gaining influence where a foundation promoting specific AI risk framings funds a staff position within a senator's office − senators operate like individual policy institutes with 20-50 personal staff who draft legislation and advise on positions. The same senator later becomes one of four key decision-makers for all US federal AI policy. This would be equivalent to a foundation placing an advisor with an MEP who then gets selected by the European Parliament President to chair the committee writing EU-wide AI regulation, but with less transparency than EU lobbying rules typically require.

[18] See AlgorithmWatch "Large Language Models Continue To Be Unreliable Concerning Elections," 2024, https://algorithmwatch.org/en/llms_state_elections/ and "ChatGPT and Co: Are AI-driven search engines a threat to democratic elections?," 2023, https://algorithmwatch.org/en/bing-chat-election-2023/

[19] Raluca Besliu, Aniket Narawad and Anna Toniolo for AlgorithmWatch, "Infrastructure or Intrusion? Europe's Conflicted Data Center Expansion," 2025 https://algorithmwatch.org/en/infrastructure-intrusion-conflict-data-center/.

[20] Michael Bird and Nathan Schepers for AlgorithmWatch, 2025, https://algorithmwatch.org/en/ai-revolution-exploitation-gig-workers/.

[21] "Report Algorithmic Discrimination", AlgorithmWatch, 2025, https://algorithmwatch.org/en/report-algorithmic-discrimination/.