You might think that people concerned about catastrophic risk from AI would generally be opposed to technology. This is not the case.

Many of these people are long-time techno-optimists and even identify with transhumanism.[1] The core claim is not "stop technology," but "let's not build something that would likely wipe out humanity." AI safety advocates often support beneficial, well-understood technologies (e.g., nuclear power, vaccines, self-driving systems, and many current uses of AI[2]), while urging strong safeguards or limits on unusually dangerous areas (e.g., on gain-of-function research, on some effects of current AI,[3] and on future, highly-capable AI). Most also acknowledge significant present-day AI benefits and potential future gains; the worry is that a takeover by future AI would forfeit all those benefits at once.

Views differ on how to reduce risk from advanced AI.

- Some propose pausing or sharply slowing frontier model development — ranging from "indefinitely" to "for a generation" to "until specific safety criteria are met" — but these proposals target frontier AI, not technology writ large.
- Some favor *differential acceleration* ("d/acc"): speeding up development of tools that constrain and secure powerful systems (e.g., evals, interpretability, sandboxing, compute governance, and defensive technologies), including using narrow AI to help make broader AI safer.
- Many safety-minded researchers inside frontier AI companies argue for continued but cautious progress to better understand and control the systems being built.

Across these strands, there is support for targeted policies: rigorous capability evaluations, red-teaming, staged deployment, liability for "foom", and restrictions on especially hazardous uses. This contrasts with some pro-technology groups (e.g., most e/accs, some open-source advocates, and many venture capitalists) who oppose almost any restriction. Nonetheless, AI safety advocates are broadly more "pro-technology" than the average person.

As an analogy, the ideal world of people concerned about bridge safety is not one with no bridges, but one with no unsafe bridges. In the same way, for people who are concerned about AI safety, the ideal world is not one with no AI, but one with no powerful unsafe AI.

---

[1] Examples include Eliezer Yudkowsky and Nick Bostrom.
[2] People who are concerned about future AI are often power-users of current AI. As an analogy, early humans might be wary of burning down an entire wooded area while enthusiastically using small fires to cook food.
[3] Salient examples include AI-induced psychosis as well as AI persuasion, but many people concerned by existential risk are also concerned by other harms from AI.

Alternative phrasings

- Are people concerned about AI safety Luddites?

Related

- How do we stop regulations from preventing the development of beneficial technol…
- How successfully have institutions managed risks from novel technology in the past?
- Would a slowdown in AI capabilities development decrease existential risk?

Scratchpad

Algon: We should mention d/acc in there somewhere.
Ability to pause, and PauseAI as a whole may be worth mentioning. Some of those people are "anti-technology" in the sense that they approve of that vibe. But PauseAI was built to be as inclusive a coalition as possible. They contain multitudes.

<Allan's draft>

In the exciting and fast-moving world of artificial intelligence, it's easy to get caught up in the excitement of technological progress and overlook the importance of safety considerations. Those concerned about existential risks are often portrayed as being inherently opposed to innovation. This characterization likely stems from the assumption that safety measures act as roadblocks to progress, and it's further fueled by the fact that some voices within the AI safety community express more extreme views, such as calling for a complete halt to AI development.

However, this portrayal is untrue for many working in the field. While opinions on balancing progress and safety vary widely, a significant portion of people dedicated to mitigating AI existential risks (x-risks) are also actively pushing innovation forward. Their goal is to ensure that AI progress is made responsibly and safely, not to hinder it.

This article aims to explore the intricate relationship between AI safety and technological progress, illustrating how these two seemingly opposing forces can work together to create a future where advanced AI systems benefit humanity while minimizing potential risks.

AI safety isn’t about hitting the brakes on technology

Many AI safety researchers are at the forefront of AI development and are deeply excited about this technology's potential. They aim to ensure we develop increasingly powerful AI systems responsibly and safely, akin to developing safety protocols for any groundbreaking technology.

Consider how safety considerations are integral to fields like aerospace or civil engineering. When engineers design a bridge, they focus on more than just making it span a great distance. They also consider factors like wind resistance, weight limits, and earthquake safety. Does this slow down bridge construction? Yes, to some extent. These calculations and precautions do add time and complexity to the process. However, this slowdown is a crucial investment: it ensures that the bridges we build are reliable and durable, preventing catastrophic failures that could cost lives and set back the entire field. While these safety measures don't necessarily accelerate progress, they do something equally valuable: make progress sustainable. They prevent disasters that could be catastrophic and lead to even longer setbacks. They create a stable foundation upon which further innovations can be built. Thus, these safety considerations are integrated into the development of bridges rather than being seen as separate or opposed to progress.

This same principle can be applied to AI development. To create a powerful AI system that reliably follows human intentions and values, you need to solve many of the core challenges in AI development. In other words, if you know how to build safe, highly capable AI, you inherently know how to make highly capable AI. So at its core AI safety research isn't separate from or opposed to advancing AI capabilities – it's an integral part of AI progress.

Diverse opinions within the AI community

Within the AI safety community, there is a wide range of perspectives on approaching the challenges and opportunities presented by advanced AI systems. Prominent figures like Yoshua Bengio, Nick Bostrom, and Eliezer Yudkowsky have contributed to the pro-technology Transhumanist movement, all of whom have contributed significantly to the pro-technology Transhumanist movement, now hold differing views on the balance between AI development and safety.

Bengio, while continuing to advance AI capabilities, emphasizes the need for robust safety measures and ethical considerations. Bostrom, who initially raised concerns about the risks of advanced AI, now cautions against overly restrictive measures that might lead to less safe AI systems being developed first. Yudkowsky, on the other hand, has become increasingly concerned about the risks of unaligned AI and advocates for a significant slowdown in AI development to address fundamental safety problems.

These diverse perspectives reflect the uncertainty surrounding AI progress, stemming from the rapid pace of advancements and the unpredictable nature of technological breakthroughs. Even top minds and experts in AI x-risk find it difficult to reach a consensus on the exact path to

follow, highlighting the complexity of balancing rapid technological AI advancement with potential risks. Navigating this uncertain terrain requires effective dialogue and collaboration among researchers, policymakers, and other stakeholders to find common ground and develop strategies that balance the need for innovation with the imperative of mitigating existential risks.

Speed bumps or full steam ahead?

An important question in the AI safety debate is whether slowing down the development of AI capabilities could truly decrease existential risk. Proponents of slowing down argue that it would give us more time to develop robust safety measures and better understand the implications of advanced AI. On the other hand, critics contend that slowing progress could lead to less safe AI systems being developed first or cause us to miss out on potential benefits that could help mitigate other existential risks. The answer is not a simple "yes" or "no," but a multifaceted consideration of various factors and potential outcomes as discussed in this article.

In this ongoing discussion, Ethereum co-founder Vitalik Buterin has proposed an approach called d/acc (defensive/decentralized/differential acceleration) that offers a nuanced perspective. This aligns with many AI safety goals while maintaining a pro-innovation stance. Some key aspects of d/acc include:

1. **Selective acceleration:** Intentionally promoting technologies that enhance human capabilities and improve our defensive posture against potential risks.
2. **Focus on human-AI collaboration:** Emphasizing technologies that augment human cognition and decision-making rather than replacing human agency.
3. **Decentralization:** Favoring decentralized development and governance of AI systems to address concerns about concentrated power.
4. **Defensive technologies:** Prioritizing the development of protective technologies, including improved cybersecurity and information verification tools.

The d/acc perspective suggests that by being intentional about the direction of our technological progress, we can reap the benefits of advanced AI while mitigating potential risks. It reinforces the idea that responsible innovation and safety considerations are not mutually exclusive, demonstrating how safety can be integrated into the core of technological advancement. This approach aligns with the views of many in the AI x-risk community who believe that responsible innovation and safety considerations can and should be integrated into the development process. By being intentional about the direction of technological progress and investing in safety-enhancing technologies and practices, we can better control the growth of AI while balancing out the risks from it.

Conclusion: Integrating Safety and Progress

So are people concerned about AI x-risk anti-progress? Far from it. It's about ensuring that our progress in AI aligns with our broader goals as a species. It's about asking the hard questions now before we find ourselves in a situation where it's too late to change course.

What would it look like if the AI safety movement succeeds? It wouldn't necessarily mean a world without advanced AI. In an ideal scenario, it would be a world where AI significantly enhances human capabilities without supplanting human agency in critical decisions. Where AI systems are sufficiently transparent that we can meaningfully scrutinize and question their operations and outputs. Where the benefits of AI are distributed more equitably, rather than exacerbating existing power imbalances.

As we stand on the brink of an AI revolution, embracing innovation and safety isn't just a good idea — it's essential. The AI safety movement doesn't promise easy solutions or certain outcomes. Instead, it advocates for a thoughtful, measured approach to one of the most powerful and potentially transformative technologies humanity has ever developed. In navigating these uncharted waters, our caution and foresight today may well determine the course of humanity's future.

Explore topic later:

Can Safety efforts help make AIs more capable?

A common misconception is that we have to choose between rapid AI advancement and excessive caution. In reality, safety and progress are complementary, not contradictory.

The relationship between AI safety and progress is sometimes mischaracterized as a tug-of-war, when it's more accurately described as a symbiotic partnership. Safety research isn't just about preventing disasters; it's actively driving innovation and enabling more advanced AI systems. Here's how:

1. **Enabling "Better" Models:** Take the development of large language models like GPT-4, Claude, and their predecessors. Researchers didn't just focus on making these models bigger and more powerful. They also implemented proactive measures like constitutional AI and reinforcement learning from human feedback (RLHF). These techniques are designed to guide AI outputs and behaviors towards greater safety and accuracy from the ground up, but they also enhance the models' reasoning abilities and overall aptness.
2. **Improving Model Reliability:** Safety research has led to breakthroughs in making AI systems more reliable and consistent. Techniques developed to prevent harmful outputs,

like content filtering and toxicity reduction, have also made models more dependable for a wider range of applications. This increased reliability is crucial for deploying AI in critical real-world scenarios.

3. **Expanding AI Applications:** By making AI systems safer and more controllable, safety research is actually expanding the domains where AI can be applied. Many sensitive areas, like healthcare or financial services, require a high degree of safety and reliability before AI can be deployed. Safety advancements are opening these doors.
4. **Addressing Scaling Challenges:** As AI systems become more powerful, new challenges emerge. Safety research is at the forefront of addressing these scaling challenges. For instance, work on AI alignment - ensuring AI systems pursue the intended goals - becomes increasingly crucial as models become more capable. This research is paving the way for the responsible development of more advanced AI.
5. **Interdisciplinary Advancements:** AI safety research often requires interdisciplinary approaches, combining insights from computer science, ethics, psychology, and other fields. This cross-pollination of ideas often leads to novel approaches and breakthroughs that benefit AI development as a whole.

Is an early focus on AI safety hindering progress?

While the push for AI safety is compelling, some critics argue that the current focus on potential risks may be premature or counterproductive to innovation. They contend that current AI capabilities are often exaggerated, and we're still far from artificial general intelligence (AGI) or superintelligence. Moreover, critics warn against the "solutionism" trap — the belief that every potential problem must have a technological fix — arguing that proposed solutions might create new, unforeseen issues.

Proponents of AI safety research argue that the potential risks of AI are far from exaggerated. They point to several factors that necessitate immediate action:

1. **Exponential Progress:** AI capabilities are advancing at an unprecedented rate. What seems like science fiction today could become reality sooner than we expect. By the time obvious dangers manifest, it may be too late to implement effective safeguards.
2. **Irreversibility:** Once highly advanced AI systems are deployed, it may be challenging or impossible to "put the genie back in the bottle." Establishing robust safety protocols now can prevent potentially irreversible consequences. Even confining such advanced AI systems seems like an uphill battle.
3. **The Alignment Problem**: To put it plainly, ensuring AI systems align with human values is extremely difficult. Starting early allows us to develop and refine alignment techniques alongside AI advancements.
4. **Avoiding the "Solutionism" Trap:** AI safety advocates acknowledge the complexity of the challenge. They're not proposing simple technological fixes, but a comprehensive approach that includes technical research, policy development, and ethical

considerations. This long-term strategy aims to create a framework for responsible AI development that can adapt to even unforeseen challenges.

5. **Economic and Social Impact:** AI is already influencing various sectors of society. Implementing safety measures now can help mitigate potential negative impacts on employment, privacy, and social structures.

The rapid pace of AI development makes it critical to address potential risks early. The argument that we're far from AGI or superintelligence doesn't negate the need for safety measures; rather, it underscores the importance of developing these safeguards alongside AI capabilities. Just as we don't wait for a bridge to collapse before implementing safety standards in engineering, we shouldn't wait for AGI to emerge before considering its implications.

Moreover, AI safety research isn't just about averting catastrophic scenarios. We're already seeing unanticipated consequences from relatively simple AI systems, from biased decision-making in hiring processes to the spread of misinformation through social media algorithms. These real-world examples highlight the need for proactive safety measures, even without AGI. The call to pause or slow down AI development isn't about stifling progress; it's about ensuring that progress is sustainable, safe, and beneficial to humanity. By addressing these concerns now, we can foster an environment where AI can flourish responsibly, maximizing its potential benefits while minimizing risks.

The relationship between safety considerations and innovation in AI is complex. While some argue that a focus on safety might slow down certain avenues of research, others contend that safety considerations can coexist with and even complement innovation. Many researchers are working to advance AI capabilities while simultaneously addressing safety concerns, suggesting that these two aspects of AI development are not mutually exclusive. The challenge lies in finding a balance that allows for continued progress while implementing necessary safeguards. This balanced approach aims to create AI systems that are not only advanced but also reliable and aligned with human values.