UNIVERSITY OF CAMBRIDGE | CASP Cambridge Programme on AI Science & Policy

Executive Summary

What if automating AI R&D triggers an intelligence explosion?

This summary and its accompanying paper mark the first research collaboration across leading academics, senior scientists of frontier AI companies, and independent experts from civil society to assess the possibility of an intelligence explosion and recommend policies to prepare for it. The effort was initiated and led by academic and civil society researchers, and its co-authors include: Turing Award winners Geoffrey Hinton (also a Nobel laureate), Yoshua Bengio, and Andrew Barto; industry leaders from OpenAI (Jakub Pachocki, Chief Scientist), Microsoft (Eric Horvitz, Chief Scientific Officer), and Anthropic (Jack Clark, Co-Founder); and researchers from numerous leading universities.

AI now plays a major role in building AI. Preliminary evidence suggests this could radically accelerate AI progress in an "intelligence explosion," where years of AI progress are compressed into months or less. Such an acceleration could be the most consequential technological development in history. Policymakers, including heads of government, urgently need to understand and prepare for this possibility.

AI is rapidly automating AI research and development (R&D). As of August 2026, Anthropic reported AI systems completing 26% of internal AI R&D work with only high-level supervision, compared to just 1% five months earlier. As of September 2026, OpenAI reported systems routinely completing R&D tasks that would take days for staff. Based on benchmark trends and other evidence, some experts think that AI R&D could be fully automated within the next few years.

Automating AI R&D could lead to an "intelligence explosion," compressing years of AI progress into months or less. Today, only thousands of researchers work on frontier AI R&D. Because AI systems can be copied and run in parallel, automating this work could add the equivalent of millions more. As these AI systems help build better successors, this effective workforce could grow even further. Preliminary evidence suggests that this feedback loop could radically accelerate AI progress, overcoming frictions such as diminishing returns to research labor and hard-to-automate tasks.

An intelligence explosion could dramatically bring forward AI's benefits, but also pose extreme risks. We might rapidly obtain superintelligent AI systems that transform many sectors of society, bringing unprecedented breakthroughs in science and industry. At the same time, these developments could arrive much faster than society can adapt to the resulting risks, such as AI-enabled pandemics, cyber attacks on critical infrastructure, and large-scale labor disruption. Furthermore, humans may have little to no oversight over automated AI R&D, raising the risk that superintelligent systems irreversibly escape human control and act to marginalize humanity. Finally, an intelligence explosion could transform modest AI capability leads into decisive ones: even the prospect of a decisive lead could provoke international conflict, and an actual one could severely erode checks and balances within and between states, companies, and branches of government.

Policymakers should urgently prioritize:

(1) **Obtaining visibility into AI R&D automation,** such as by supervising frontier AI companies through embedded auditors and requiring these companies to report key indicators, including the extent of AI R&D automation, the pace of AI progress, how R&D spending is allocated, and whether and how AI systems are used in high-stakes R&D decisions.

(2) **Developing ways to steer and constrain an intelligence explosion,** such as by setting concrete safety requirements for continued development and deployment, setting a speed limit on capabilities growth, monitoring high-stakes AI R&D experiments and building the option to shut them down if needed, requiring that AI R&D take place in secure and isolated environments (e.g., air-gapped networks), establishing international incident-sharing, clarifying how deterrence would apply around an intelligence explosion so as to avoid unexpected escalation, securing international agreements to pace progress if needed without fear of falling behind, and developing verification tools in advance for such agreements.

(3) **Preparing society for the impacts of an intelligence explosion,** such as by ensuring that AI follows the law; providing citizens, civil society, and branches of government the capabilities to detect, document, and contest unlawful or harmful uses of AI; and creating emergency response plans for major AI-driven cybersecurity or biosecurity incidents, labor market disruption, geopolitical instability, and loss of control over highly capable AI systems.

This Executive Summary is provided by the Cambridge Programme on AI Science & Policy and is written by a subset of the authors.