Gladstone AI

Defense in Depth: An Action Plan to Increase the Safety and Security of Advanced AI

This publication was produced for review by the United States Department of State. It was prepared by Gladstone February 26, 2024 and the contents are the responsibility of the authors. The authors' views expressed in this publication do not reflect the views of the United States Department of State or the United States Government.

Organization Gladstone AI Inc. (hello@gladstone.ai)

Authors

Edouard Harris*

Jeremie Harris

Mark Beall

* Lead and corresponding author.
Contact: edouard@gladstone.ai

Executive summary

The recent explosion of progress in advanced artificial intelligence (AI) has brought great opportunities, but it is also creating entirely new categories of weapons of mass destruction-like (WMD-like) and WMD-enabling catastrophic risks[1] [1–4]. A key driver of these risks is an acute competitive dynamic among the frontier AI labs[2] that are building the world's most advanced AI systems. All of these labs have openly declared an intent or expectation to achieve human-level and superhuman artificial general intelligence (AGI)[3] — a transformative technology with profound implications for democratic governance and global security — by the end of this decade or earlier [5–10].

The risks associated with these developments are global in scope, have deeply technical origins, and are evolving quickly. As a result, policymakers face a diminishing opportunity to introduce technically informed safeguards that can balance these considerations and ensure advanced AI is developed and adopted responsibly. These safeguards are essential to address the critical national security gaps that are rapidly emerging as this technology progresses.

Frontier lab executives and staff have publicly acknowledged these dangers [11–13]. Nonetheless, competitive pressures continue to push them to accelerate their investments in AI capabilities at the expense of safety and security. The prospect of inadequate security at frontier AI labs raises the risk that the world's most advanced AI systems could be stolen from their U.S. developers, and then weaponized against U.S. interests [9]. Frontier AI labs also take seriously the possibility that they could at some

[1] By **catastrophic risks**, we mean risks of catastrophic events up to and including events that would lead to human extinction.

[2] By **frontier AI labs**, we mean the organizations that are involved in building cutting-edge, general-purpose AI systems, and whose research programs are explicitly aimed at, or could plausibly lead to, the development of artificial general intelligence or AGI. Examples include OpenAI, Google DeepMind, and Anthropic.

[3] By **AGI**, we mean an AI system that can outperform humans across all economic and strategically relevant domains, such as producing practical long-term plans that are likely to work under real world conditions.

point lose control[4] of the AI systems they themselves are developing [5,14], with potentially devastating consequences to global security.

Given the growing risk to national security posed by rapidly expanding AI capabilities from weaponization and loss of control — and particularly, the fact that the ongoing proliferation of these capabilities serves to amplify both risks — there is a clear and urgent need for the U.S. government to intervene.

This action plan is a blueprint for that intervention. Its aim is to **increase the safety and security of advanced AI by countering catastrophic national security risks from AI weaponization and loss of control**. It was developed over thirteen months, and informed by conversations with over two hundred stakeholders from across the U.S., U.K., and Canadian governments; major cloud providers; AI safety organizations; security and computing experts; and formal and informal contacts at the frontier AI labs themselves. The actions we propose follow a sequence that:

- Begins by establishing interim safeguards to stabilize advanced AI development, including export controls on the advanced AI supply chain;

- Leverages the time gained to develop basic regulatory oversight and strengthen U.S. government capacity for later stages;

- Transitions into a domestic legal regime of responsible AI development and adoption, safeguarded by a new U.S. regulatory agency; and

- Extends that regime to the multilateral and international domains.

[4] **Loss of control** due to **AGI alignment failure** is a potential failure mode under which a future AI system could become so capable that it escapes all human efforts to contain its impact.

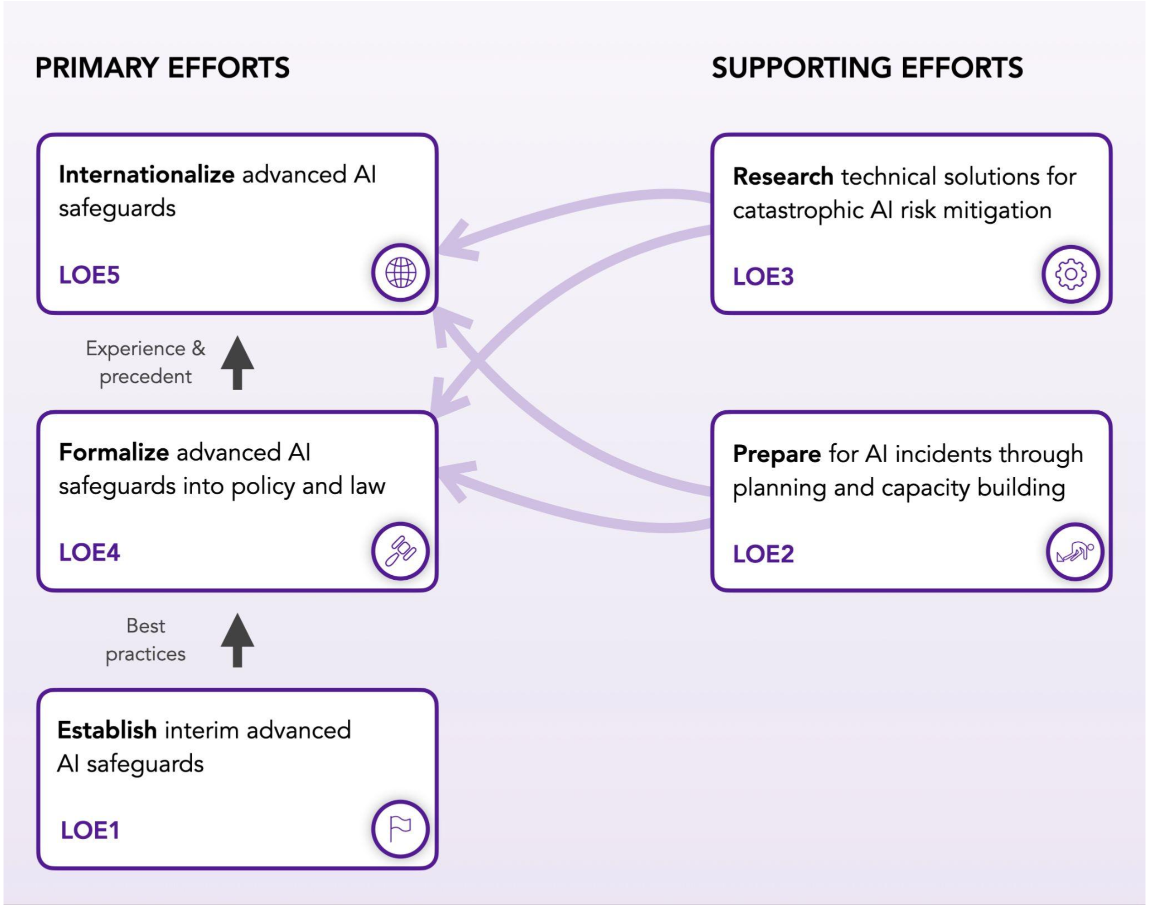

PRIMARY EFFORTS
SUPPORTING EFFORTS
Internationalize advanced AI safeguards
LOE5
Experience & precedent
Formalize advanced AI safeguards into policy and law
LOE4
Best practices
Establish interim advanced AI safeguards
LOE1
Research technical solutions for catastrophic AI risk mitigation
LOE3
Prepare for AI incidents through planning and capacity building
LOE2

Figure 1. Overview of the action plan and its component LOEs.

The U.S. government and its allies and partners, in close partnership with industry, can achieve this aim by implementing five mutually supporting lines of effort (LOEs). These LOEs will **establish** (LOE1), **formalize** (LOE4), and **internationalize** (LOE5) safeguards on advanced AI development, while **increasing preparedness** (LOE2) and **building technical capacity and capability** (LOE3). Some of the measures we propose are unprecedented, but after consulting with stakeholders and experts, we believe they are proportionate to the magnitude and urgency of the risk we face.

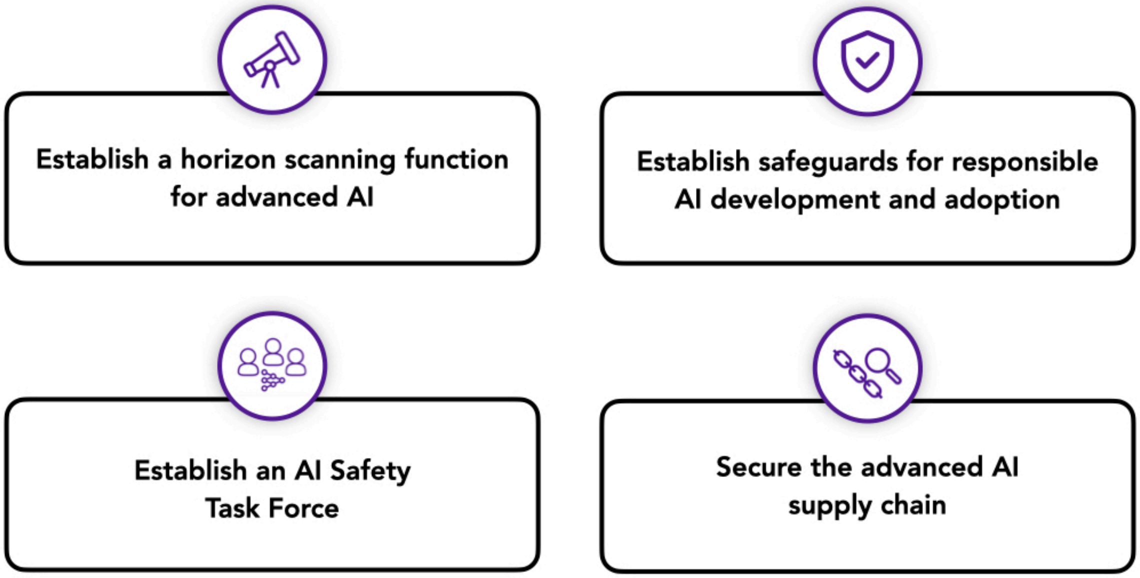
Because of the severity, uncertainty, and irreversibility of these risks, an action plan to address them needs to offer a wide margin of safety. This plan follows the principle of **defense in depth**, in which multiple overlapping controls combine to offer resilience against any single point of failure. We frame tradeoffs in terms of AI breakout timelines, the amount of time it would take an actor to train an AI system from scratch to equal the current state-of-the art under various expert-vetted assumptions. And we illustrate this framework with an example regulatory regime that targets an AI breakout timeline of 18 months to train a GPT-4 equivalent AI model under worst-case assumptions

(LOE4). We expect regulators to set their own thresholds and update them depending on the lead times required by contingency planners (LOE2), and in response to future technological developments.

AI development and governance is complicated and dynamic, and exists at the intersection of multiple unsolved questions in engineering, policy, and fundamental research. As a result, some of our recommendations may be flawed and should be vetted by relevant subject-matter experts. Nonetheless, we believe that this action plan is the most complete framework proposed so far to support an informed, effective, and rapid response to the emerging threats we face at this historic inflection point.

We include a brief summary of each of the plan's LOEs below.

LOE1: Establish interim safeguards to stabilize advanced AI development



Establish a horizon scanning function for advanced AI

Establish safeguards for responsible AI development and adoption

Establish an AI Safety Task Force

Secure the advanced AI supply chain

Current frontier AI development poses urgent and growing risks to national security. As components of the AI supply chain proliferate, these risks will become increasingly challenging to contain. Moreover, the pace of development in AI is now so rapid that an ordinary policymaking process could be overtaken by events by the time the resulting policies take effect.

This LOE outlines possible actions the Executive Branch could take to **buy down catastrophic AI risk in the near term (1-3 years), while setting the conditions for successful long-term AI safeguards**. These actions are:

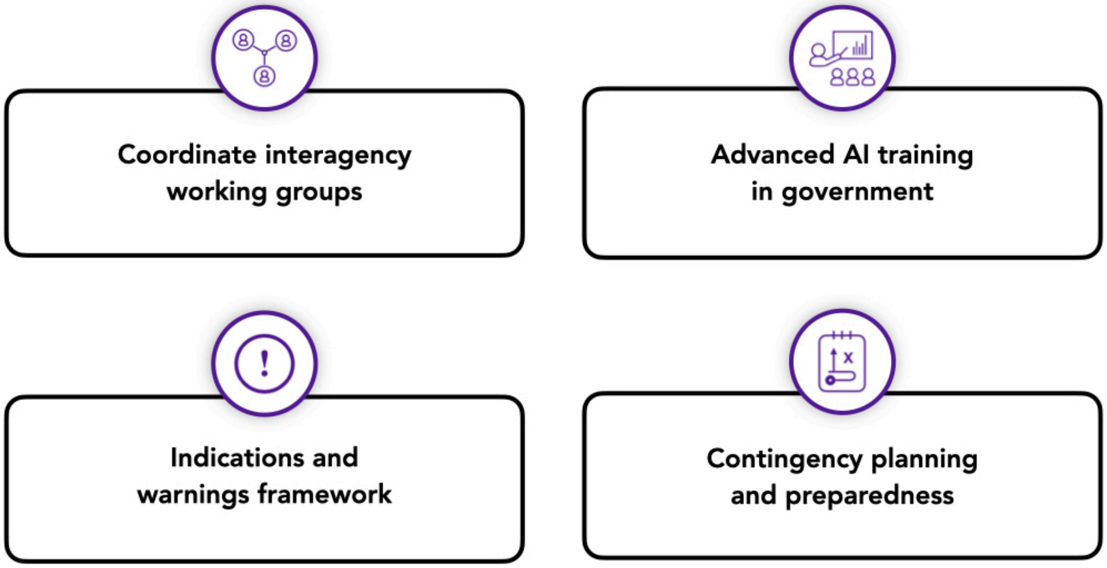
- Creating an AI Observatory (AIO) to monitor developments in advanced AI and ensure that the U.S. government's view of the field is up-to-date and reliable;

- Mandating an interim set of responsible AI development and adoption (RADA) safeguards for advanced AI systems and their developers;

- Creating an interagency AI Safety Task Force (ASTF) to coordinate implementation and oversight of RADA safeguards; and

- Putting in place a set of controls on the advanced AI supply chain calibrated to preserve U.S. government flexibility in the face of unpredictable risks.

LOE2: Strengthen capability and capacity for advanced AI preparedness and response



Coordinate interagency working groups

Advanced AI training in government

Indications and warnings framework

Contingency planning and preparedness

Advanced AI and AGI risk mitigation will engage a broad set of U.S. government equities. However, understanding of the advanced AI landscape is uneven. Mitigation measures require advance planning, coordination, and a broad understanding of risk signals to be most successful, which entails substantial capacity-building.

This LOE outlines specific actions that the U.S. government could take to **increase its preparedness for rapidly addressing incidents related to advanced AI and AGI development and deployment**. These actions are:

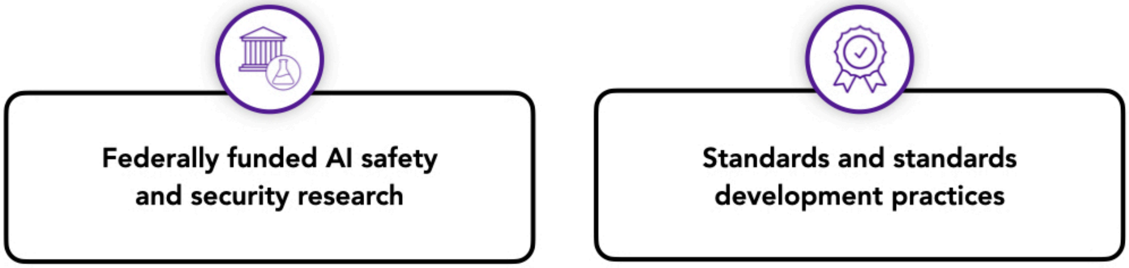
- Directing the establishment of interagency working groups for the LOEs listed in this action plan;

- Increasing preparedness and response capacity and capability through education and training;

- Coordinating the development of an Indications and Warnings (I&W) framework for advanced AI and AGI incidents; and

- Coordinating the development of scenario-based contingency plans.

LOE3: Increase national investment in technical AI safety research and standards development



Federally funded AI safety and security research

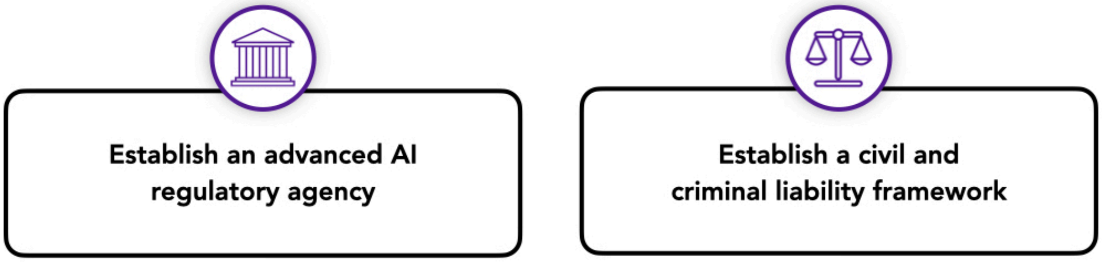
Standards and standards development practices

The acceleration of investment in AI capabilities is outpacing the development of proportionate technical safeguards against advanced AI and AGI risks [5]. If this continues, frontier AI labs may find themselves unable to meet the safety and security challenges posed by their own systems. Unless strong technical safeguards are designed, standardized, and broadly applied, continued development and adoption of frontier AI systems could create significant risks.

This LOE outlines specific actions the U.S. government could take to **strengthen domestic technical capacity in advanced AI safety and security, AGI alignment, and other technical AI safeguards.** These actions include:

- Directly funding advanced AI safety and security research including AGI-scalable alignment research; and
- Developing, regularly reviewing, and promulgating safety and security standards for responsible AI development and adoption.

LOE4: Formalize safeguards for responsible AI development and adoption by establishing an AI regulatory agency and legal liability framework



Establish an advanced AI regulatory agency

Establish a civil and criminal liability framework

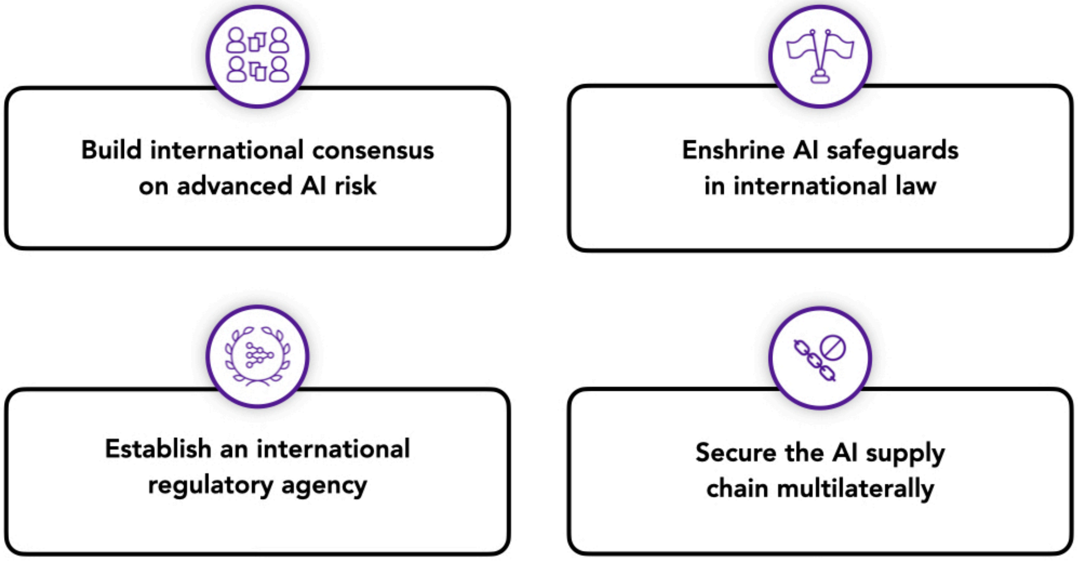
Interim regulations may be insufficient to address the unique risks and challenges of advanced AI. A legal framework for AI regulation and liability, that directly addresses catastrophic risk through detailed and flexible responsible AI development and adoption (RADA) safeguards, is essential to promote long-term stability and cover any gaps in existing authorities.

This LOE outlines specific actions the Legislative Branch could take to **establish the conditions for long-term (4+ years) domestic AI safety and security**. These actions include:

- Creating a Frontier AI Systems Administration (FAISA), a regulatory agency with rulemaking and licensing powers to oversee AI development and deployment, consistent with a set of RADA safeguards derived from contingency planning requirements; and

- Establishing a criminal and civil liability regime that could include defining responsibility for AI-induced damages; determining the extent of culpability for AI accidents and weaponization across all levels of the AI supply chain; and defining emergency powers to respond to dangerous and fast-moving AI-related incidents which could cause irreversible national security harms.

LOE5: Enshrine AI safeguards in international law and secure the AI supply chain



Build international consensus on advanced AI risk

Enshrine AI safeguards in international law

Establish an international regulatory agency

Secure the AI supply chain multilaterally

The rise of advanced AI and AGI has the potential to destabilize global security in ways reminiscent of the introduction of nuclear weapons. As advanced AI matures and the elements of the AI supply chain continue to proliferate, countries may race to acquire the resources to build sovereign advanced AI capabilities. Unless carefully managed, these competitive dynamics risk triggering an AGI arms race and increase the likelihood of global- and WMD-scale fatal accidents, interstate conflict, and escalation.

This LOE outlines near-term diplomatic actions and longer-term measures the U.S. government could take to **establish an effective AI safeguards regime in international law while securing the AI supply chain**. These actions include:

- Building a domestic and international consensus on catastrophic AI risks and necessary safeguards;
- Enshrining those safeguards in international law;
- Establishing an International AI Agency (IAIA) to monitor and verify adherence to those safeguards; and
- Establishing an AI Supply Chain Control Regime (ASCCR) with allies and partners to limit the proliferation of advanced AI technologies.

The specific recommendations in each of these LOEs are semi-flexible. In some cases, functions for which we recommend establishing a new task force or agency (e.g. LOE5) could be incorporated into existing or recently established U.S. government offices, systems, or entities.

Several of these LOEs also call for bold action beyond what has been required in previous periods of rapid technological change. We do not make these recommendations lightly. Rather, they reflect the unprecedented challenge posed by rapidly advancing AI capabilities which create the potential for catastrophic risks fundamentally unlike any that have previously been faced by the United States. They also reflect a multitude of unique challenges that make the threats resistant to single-approach solutions. These include:

- The severity of worst case outcomes is extreme;
- The timescale and degree of risk are highly uncertain;
- The entities developing frontier AI systems are incentivized to invest in capabilities at the expense of safety and security;
- The advanced AI supply chain is especially prone to proliferation, particularly in the case of open-access AI models;
- The geopolitical landscape may pose a further challenge to coordination; and
- The introduction of excessive regulation in this domain may harm innovation and competitiveness.

To paraphrase a safety researcher at a frontier lab, the risk from this technology will be at its most acute just as it seems poised to deliver its greatest benefits. Given these factors, inaction is likely to erode decisionmaker flexibility and narrow options in the face of a rapidly evolving risk landscape. But by taking bold action, the United States can seize a unique opportunity to lead the domestic, scientific, and international efforts that will meet the needs of this historic moment.

Bibliography

[1] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Risks". (2023).

[2] Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2019). Optimal policies tend to seek power. In *arXiv [cs.AI]*. http://arxiv.org/abs/1912.01683

[3] Clark, J. [jackclarkSF]. (2022, August 7). *Malware is bad now but will be extremely bad in the future due to intersection of RL + code models + ransomware economic incentives. That train is probably 1-2 years away based on lag of open source replication of existing private models, but it's on the tracks*. Twitter. https://twitter.com/jackclarkSF/status/1556181432522797056

[4] Amodei, D. (2023, July 25). *Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic For a hearing on "Oversight of A.I.: Principles for Regulation" Before the Judiciary Committee Subcommittee on Privacy, Technology, and the Law United States Senate*. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf

[5] Leike, J., & Sutskever, I. (2023, July 5). *Introducing superalignment*. Openai.com. https://openai.com/blog/introducing-superalignment

[6] Altman, S. (2023, February 24). *Planning for AGI and beyond*. Openai.com. https://openai.com/blog/planning-for-agi-and-beyond

[7] *Core Views on AI Safety: When, Why, What, and How*. (2023, March 8). Anthropic.com. https://www.anthropic.com/index/core-views-on-ai-safety

[8] Bove, T. (2023, May 3). *CEO of Google's DeepMind says we could be 'just a few years' from A.I. that has human-level intelligence*. Yahoo Finance. https://finance.yahoo.com/news/ceo-google-deepmind-says-could-213237542.html

[9] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Notable players, their products and capabilities". (2023).

[10] Wiggers, K., Coldewey, D., & Singh, M. (2023, April 6). Anthropic's $5B, 4-year plan to take on OpenAI. *TechCrunch*. https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/

[11] Perrigo, B. (2023, January 12). DeepMind's CEO helped take AI mainstream. Now he's urging caution. *Time*. https://time.com/6246119/demis-hassabis-deepmind-interview/

[12] Perrigo, B. (2023, May 30). AI is as risky as pandemics and nuclear war, top CEOs say, urging global cooperation. *Time*. https://time.com/6283386/ai-risk-openai-deepmind-letter/

[13] Bove, T. (2023, May 30). *Sam Altman and other technologists warn that A.I. poses a 'risk of extinction' on par with pandemics and nuclear warfare*. Fortune. https://fortune.com/2023/05/30/sam-altman-ai-risk-of-extinction-pandemics-nuclear-warfare/

[14] *Anthropic's Responsible Scaling Policy*. (2023, September 19). Anthropic. https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf