HOW DOES ACCESS IMPACT RISK?

ASSESSING AI FOUNDATION MODEL RISK ALONG A GRADIENT OF ACCESS

ZOË BRAMMER

DECEMBER 2023

IST Institute for SECURITY + TECHNOLOGY

How Does Access Impact Risk?
Assessing AI Foundation Model Risk Along a Gradient of Access

December 2023
Author: Zoë Brammer
Contributors: Anthony Aguirre, Markus Anderljung, Chloe Autio, Ramsay Brown, Chris Byrd, Gaia Dempsey, David Evan Harris, Vaishnavi J., Landon Klein, Sébastien Krier, Jeffrey Ladish, Nathan Lambert, Aviv Ovadya, Elizabeth Seger, Deger Turan
Design: Sophia Mauro

IST may provide information about third-party products or services, including security tools, videos, templates, guides, and other resources included in our cybersecurity toolkits (collectively, “Third-Party Content”). You are solely responsible for your use of Third-Party Content, and you must ensure that your use of Third-Party Content complies with all applicable laws, including applicable laws of your jurisdiction and applicable U.S. export compliance laws.

About the Institute for Security and Technology

The Institute for Security and Technology (IST) is a global non-profit, non-partisan think tank uniquely situated in Silicon Valley with deep ties to Washington, D.C. and other global capitals.

As new technologies present humanity with unprecedented capabilities, they can also pose unimagined risks to global security. IST's mission is to bridge gaps between technology and policy leaders to collaboratively solve these emerging security challenges *together*. We have the access and relationships to unite the right experts via bespoke convening mechanisms, with agility, and at the right time. We are translators, conveners, and communicators who leverage unique problem-solving approaches to tackle some of the world's toughest emerging security threats.

Our portfolio is organized across three analytical pillars: the **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society; and the **Future of Digital Security**, examining the systemic security opportunities and risks of societal dependence on digital technologies.

Acknowledgments

This work is inherently collaborative. As researchers, conveners, and facilitators, IST is immensely grateful to the members of this working group for their insights, dedication, willingness to engage in honest and healthy debate, and the time that each of them generously volunteered to this effort. By working collectively to identify where risk is arising and where it is likely to continue to intensify, their contributions are what allowed this process to bear fruit. We are also immensely grateful for the generous support of the Patrick J. McGovern Foundation, whose funding allowed us to continue this project through the lens of IST's Applied Trust and Safety program. AI is too vast a set of tools, capabilities, and communities for any one organization to manage the risks and opportunities on its own. This effort reflects the cross-sectoral, public-private efforts needed more broadly across the ecosystem to ensure AI is beneficial for us all.

While each working group member does not necessarily endorse everything written in this report, we extend our gratitude to the following contributors and editors in particular:

- Anthony Aguirre
- Markus Anderljung
- Chloe Autio
- Ramsay Brown
- Chris Byrd
- Gaia Dempsey
- David Evan Harris
- Vaishnavi J.
- Landon Klein
- Sébastien Krier
- Jeffrey Ladish
- Nathan Lambert
- Aviv Ovadya
- Elizabeth Seger
- Deger Turan

Additionally, not everyone in the working group could choose to be named openly as contributors. We are just as grateful to them, including those individuals we have worked closely with and who helped to inspire the initial effort.

Finally, the author extends her gratitude to Steve Kelly and Philip Reiner for the support, guidance, and strategy they provided in the drafting and refining of this report.

Table of Contents

Executive Summary **1**

Introduction **3**

Methodology **4**

History of Access to AI Foundation Models **4**

Categories of Opportunity and Risk **11**

Identified Opportunities 11

Identified Risks 13

Gradient of Access to AI Foundation Models **16**

Level 0: Fully Closed *17*

Level 1: Paper Publication *17*

Level 2: Query API Access *18*

Level 3: Modular API Access *19*

Level 4: Gated Downloadable Access *20*

Level 5: Non-Gated Downloadable Access *21*

Level 6: Fully Open Access *21*

Matrix: Assessing AI Foundation Model Risk Along a Gradient of Access **22**

Reading the Matrix 22

Source of Risk 23

Risk Breakdown by Category 25

Fueling a race to the bottom *25*

Malicious use *27*

Capability overhang *29*

Compliance failure *31*

Taking the human out of the loop *32*

Reinforcing bias *34*

Conclusion **36**

Looking Ahead **39**

Executive Summary

In the last few years, a number of leading AI labs, including OpenAI, Anthropic, Meta, Google Deepmind, Inflection AI, and Stability AI, among others, have released highly capable artificial intelligence (AI) foundation models.[1] While some models remain highly restricted, limiting who can access the model and its components, other models and their developers provide fully open access to model weights and architecture. The potential benefits of these more open postures are generally well understood, but specific risks much less so. As a result, this effort turned our attention to the risks, seeking to answer the question:

> ***How does access to AI foundation models and their components impact the risk they pose to individuals, groups, and society?***

This question has been asked and answered, but in broad and undefined ways. There is currently no clear mechanism for understanding the risks that may arise as models are opened to greater levels of access, not least because, in light of recent industry developments, it is clear that the "open" versus "closed" framing does not sufficiently capture the full spectrum of access in the AI ecosystem.

To address this analytical gap and identify the risks and opportunities posed by increased access to highly capable foundation models, the Institute for Security and Technology (IST) convened a working group and conducted surveys and expert interviews with representatives from leading AI labs, industry, think tanks, and academia over the course of the last six months.

This report is the result of this work, combined with our own independent research and analysis. This report first provides a brief background on the technical history and ongoing debate around open access to foundation models. It then identifies categories of opportunity and risk created by cutting-edge AI, and outlines a gradient of access to these models. Finally, this report maps the relationship between specific types of risk at varying levels of access in the form of a matrix.

[1] We draw from Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models," Centre for the Governance of AI (2023), which defines foundation models as those machine learning models that "exhibit high performance across a broad domain of cognitive tasks, often performing the tasks as well as, or better than, a human."

This novel analytical approach produces a number of preliminary conclusions regarding real and potential risks, each of which will be further explored throughout this report.

Conclusions include:

- Uninhibited access to AI foundation models and their components significantly increases the risk these models pose across a range of categories, as well as the ability for malicious actors to abuse AI capabilities and cause harm.

- Specifically, as access increases, the risk of malicious use (such as fraud and other crimes, the undermining of social cohesion and democratic processes, and/or the disruption of critical infrastructure), compliance failure, taking the human out of the loop, and capability overhang (model capabilities and aptitudes not envisioned by their developers) all increase.

- At the highest levels of access, the risk of a “race to the bottom”—a situation in which conditions in an increasingly crowded field of cutting-edge AI models might incentivize developers and leading labs to cut corners in model development—increases when assuming a “winner takes all” dynamic.

- As access increases, the risk of reinforcing bias—the potential for AI to inadvertently further entrench existing societal biases and economic inequality as a result of biased training data or algorithmic design—fluctuates.

Introduction

Allowing fully open access to powerful AI models is a new phenomenon. Most advanced general purpose language models have been built almost exclusively by a few commercial, well-funded AI labs, which tend to limit access to them via a frontend product integration for general users and an Application Programming Interface (API) for developers. While some AI components like toolkits, frameworks, and benchmark data sets have routinely been open sourced, in recent months, labs offering direct access to the source code of ever-larger foundation models have drastically altered AI ecosystem dynamics and development cycles.

Broader access to AI models presents a number of benefits; for example, increased transparency fuels rapid innovation and enables stress-testing, red teaming, and vulnerability identification by a wider developer and user base. However, increased access also widens the space for actors who might misuse models by lowering the barrier of entry for model development and fine-tuning. Open access to these models raises questions about safety and security, among a range of other issues.

In this report, we first briefly delineate a history of access to AI foundation models, exploring technical developments and public discourse in the ecosystem. We then outline in detail distinct categories of opportunity and risk posed by the development and proliferation of cutting-edge AI technologies. In an effort to define levels of access, the report next identifies a gradient of access to AI foundation models ranging from fully closed, paper publication, query API access, and modular API access to gated downloadable access, non-gated downloadable access, and fully open access. Finally, the report introduces an assessment matrix that identifies the impact of each level of access across each category of risk.

It is our intent to provide a more specific, detailed breakdown of these elements to inform the broader debate and the decisions to be made by developers, industry leaders, and policymakers to anticipate and hopefully mitigate these risks. The conclusions reached so far through the efforts of this working group point to the need for clear technical mechanisms and policy interventions in order to maintain continued benefits while ensuring these new capabilities do not cause harm. In the end, the aim of this project is to help identify, reduce, and mitigate the potential for that harm.

Methodology

This report and accompanying matrix are the result of extensive research and collaboration. The categories of risk and opportunity outlined in this report are drawn from IST staff research; expert interviews and a survey sent to representatives of leading AI labs, think tanks, governments, and academic institutions; and a series of closed-door working group meetings with AI developers, researchers, and practitioners. The report draws on existing research and working group insights to develop the gradient of access outlined in this report and the resulting matrix. The matrix was iteratively developed through many rounds of debate, discussion, and synchronous and asynchronous refinement.

Due to the rapidly evolving technological landscape and the novelty of many topics discussed in this report, we note that not all of our conclusions are derived from existing research or data. On occasion, we rely on the collective judgment of experts in our working group to address gaps in the available literature and assess the impact of what are inherently novel technical capabilities. In such cases, this report calls attention to such judgments and provides supporting rationale.

History of Access to AI Foundation Models

To effectively identify the risks and opportunities resulting from increased access to highly capable foundation models, it is valuable to first understand the history of the ecosystem, as well as its current dynamics. This section outlines similarities between past open source debates, a brief timeline of relevant major technological advances, and attempts to capture the prevailing themes accompanying public discourse.[2]

[2] Given the rapidly changing nature of the AI ecosystem, we note that some of the data and information in this section may already be out of date by the time of publication.

Open Source Framing in the AI Ecosystem

Many working group experts noted that public discourse on access to AI models tends to borrow from the decades-long debate on the risks and benefits of open source vs. proprietary systems. This includes the late 1990s debate regarding the use of open source software in supercomputing systems, and more recent discussions about open source security and reliability,[3] the cost and accessibility of widely-used software, the mechanisms for funding innovation, the enabling of software customizability, market competition and legal and compliance issues, and the desire for interoperability and open standards. For example, a September 2000 report from the President's Information Technology Advisory Committee argues that the open source development model represents a viable strategy for producing high quality software and offers potential security advantages over the traditional proprietary development model, while acknowledging that not all risks can be foreseen.[4,5]

Though today's technological frontier looks vastly different than that of the 1990s and early 2000s, discourse around the openness of AI models remains similar: while openness can enable rapid innovation and scientific progress, and even increase security and the hardening of systems over time, it can simultaneously exacerbate existing risks, or even create significant new ones. However, in light of recent industry developments, it is clear that the "open" versus "closed" framing does not sufficiently capture the full spectrum of access or risk in the AI ecosystem. Additionally, developments in the AI space have further complicated the already complex and often contentious definition of the open source term itself, with some companies claiming to have "open sourced" a model which does not adhere to the Open Source Institute's definition of open source software. While there are many open source licensing approaches,[6] most grant developers the right to freely modify and build on top of existing open source software. In contrast, Meta branded their LLaMA 2 model as open source, but requires a license fee for developers with more than 700 million daily users, and disallows other models from training on LLaMA 2.[7]

[3] President's Information Technology Advisory Committee, Report to the President: Transforming Access to Government Through Technology, accessed November 6, 2023, https://www.nitrd.gov/pubs/pitac/pres-transgov-11sep00.pdf.

[4] President's Information Technology Advisory Committee, Report to the President.

[5] Darrell M. West et al., "How Open-Source Software Shapes AI Policy," Brookings, August 10, 2021, https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy/.

[6] "Licenses," Open Source Initiative, accessed December 4, 2023, https://opensource.org/licenses/.

[7] Emilia David, "Meta's AI Research Head Wants Open Source Licensing to Change," *The Verge*, October 30, 2023, https://www.theverge.com/2023/10/30/23935587/meta-generative-ai-models-open-source.

A Timeline of Major Technological Advances

While some in the technical community have been developing open source AI models for years, EleutherAI, founded in July 2020, was the first company to coalesce around the movement. Their work focused initially on "providing access to cutting-edge AI technologies by training and releasing models, and promoting open science norms in Natural Language Processing."[8] For two years, the movement slowly gained steam, until the summer of 2022, when industry efforts to promote increased public access to AI models significantly expanded. Hugging Face, a French-American company that develops tools and resources to build, deploy, and train machine learning models, enlisted BigScience,[9] a coalition of volunteer researchers and academics (many of whom were members of EleutherAI), in its efforts to develop a fully open foundation model. The subsequent release of BLOOM,[10] a multilingual, open source system designed for researchers, marked a significant step forward in the general public's ability to access and tweak AI models. Another early move came in the fall of 2022, when Meta contributed its open source PyTorch machine learning framework to the Linux Foundation to further accelerate the development and accessibility of the technology.[11]

In February 2023, as discourse around the risks and opportunities posed by increased access to AI models became increasingly mainstream, Meta released LLaMA under a noncommercial license focused on research use cases.[12] Meta stated that "access to the model will be granted on a case-by-case basis to academic researchers; those affiliated with organizations in government, civil society, and academia; and industry research laboratories around the world."[13] Upon its release, Yann LeCun, Meta's chief AI scientist, said, "the platform that will win will be the open one."[14] One week after Meta began fielding requests to access LLaMA, the model and model weights leaked online,[15] sparking further debate about the technical mechanisms and business strategies that facilitate access to cutting-edge research in a time of unprecedented technological change.

[8] "About," EleutherAI, accessed December 1, 2023, https://www.eleuther.ai/about.

[9] "Update: Introducing the World's Largest Open Multilingual Language Model - Bloom," BigScience, July 23, 2023, https://bigscience.huggingface.co/blog/bloom.

[10] "Update," BigScience.

[11] The Linux Foundation, "Meta Transitions Pytorch to the Linux Foundation, Further Accelerating AI/ML Open Source Collaboration," press release, September 13, 2022, https://www.linuxfoundation.org/press/press-release/meta-transitions-pytorch-to-the-linux-foundation.

[12] "Introducing Llama: A Foundational, 65-Billion-Parameter Language Model," Meta (blog), February 24, 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai/.

[13] "Introducing Llama," Meta.

[14] Cade Metz and Mike Isaac, "In Battle Over A.I., Meta Decides to Give Away Its Crown Jewels," *New York Times*, May 18, 2023, https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html.

[15] @micro_charm, "Facebook Llama Is Being Openly Distributed via Torrents," *Hacker News*, March 3, 2023, https://news.ycombinator.com/item?id=35007978.

Following the release of LLaMA and the online leak, the AI ecosystem saw a period of intense and rapid innovation, with mere days separating major developments. Variants of LLaMA with a range of unique functionalities proliferated,[16] many of which built on each other. Notably, Hugging Face then teamed up with Meta for its LLaMA 2 release, allowing for its integration on the Hugging Face platform instead of further developing BLOOM.[17]

Also in February 2023, researchers at Google DeepMind open sourced their TRAnsformer Compiler for RASP (Tracr),[18,19] a tool that a Google DeepMind study suggested might be used in controlled experiments to reveal insights into the components of some large language models (LLMs).[20]

In April 2023, Anthropic quietly expanded access to the "private alpha" version of its chat service Claude at an open source AI meetup in San Francisco.[21] Today, anyone can use Claude on the chatbot client Poe,[22] and many developers have access to Claude and Claude 2 via APIs. Further, Claude 2 is accessible to U.S. and UK users via a beta chat system.[23] Similarly, Stability AI released a suite of open source LLMs called StableLM that were developed in collaboration with the nonprofit EleutherAI.[24,25] In a blog post,[26] Stability AI announced that its models are now available for developers to use and adapt on GitHub. The company made its text-to-image AI available in several ways,[27] including a public demo, a software beta, and a full download of the model, allowing developers to alter the tool and design various integrations.

In May 2023, Abu Dhabi's Advanced Technology Research Council (ATRC) open sourced the foundation model Falcon 40B for research and commercial use, marking a significant

[16] Unique functionalities include instruction tuning, quality improvements, multimodality, and Reinforcement Learning from Human Feedback (RLHF), among others.

[17] Philipp Schmid et al., "Llama 2 Is Here - Get It on Hugging Face," Hugging Face - Blog, July 18, 2023, https://huggingface.co/blog/llama2.

[18] David Lindner et al., "Tracr: Compiled transformers as a laboratory for interpretability," arXiv, updated November 3, 2023, https://arxiv.org/abs/2301.05062.

[19] Tracr translates human-readable programs into transformer models, and is intended for research in mechanistic interpretability, a form of reverse engineering.

[20] Tanushree Shenwai, "Deepmind Open Sources Tracr: A Tool for Compiling Human-Readable Code to the Weights of a Transformer Model," *MarkTechPost*, March 2, 2023, https://www.marktechpost.com/2023/03/02/deepmind-open-sources-tracr-a-tool-for-compiling-human-readable-code-to-the-weights-of-a-transformer-model/.

[21] Sharon Goldman, "OpenAI Rival Anthropic Introduces Claude, an AI Assistant to take on ChatGPT," *Venture Beat*, March 14, 2023, https://venturebeat.com/ai/google-funded-anthropic-introduces-claude-chatgpt-rival-through-chat-and-api/.

[22] Michael Nuñez, "How to Chat with OpenAI's GPT-4 and Anthropic's Claude Right Now," *Venture Beat*, March 16, 2023, https://venturebeat.com/ai/how-to-chat-with-openai-gpt-4-and-anthropic-claude-now/.

[23] Anthropic, "Claude 2," press release, July 11, 2023, https://www.anthropic.com/index/claude-2.

[24] Stability-AI/StableLM, "StableLM: Stability AI Language Models," GitHub repository, accessed November 6, 2023, https://github.com/Stability-AI/StableLM .

[25] "EleutherAI," EleutherAI, accessed November 6, 2023, https://www.eleuther.ai/.

[26] Stability AI, "Stability AI Launches the First of Its Stable LM Suite of Language Models," press release, April 19, 2023, https://stability.ai/blog/stability-ai-launches-the-first-of-its-stablelm-suite-of-language-models.

[27] Emma Roth, "Stability AI Announces New Open-Source Large Language Model," *The Verge*, April 19, 2023, https://www.theverge.com/2023/4/19/23689883/stability-ai-open-source-large-language-model-stablelm.

milestone for foundation model development and the open access push outside of the United States.

Also in May 2023, an internal Google memo was leaked online alleging the challenges that increased access to cutting-edge AI models pose to leading AI labs.[28] It stated that open source models are "lapping" Google by addressing "major open problems" facing leading labs. Some examples included LLMs designed to be used on a smartphone, such as running foundation models on a Pixel 6;[29] scalable personal AI, such as fine tuning personalized AI models on laptops;[30] and multimodality,[31] where models can integrate and respond to numerous mediums of content or data, such as translating natural language prompts into image outputs.

Also in May, Stability AI published a paper stressing the importance of open sourcing AI models for transparency and competition in advance of a congressional hearing on the topic,[32] while Meta's Yann LeCun called the growing secrecy around AI models developed by Google and OpenAI a "huge mistake," arguing that consumers and governments will refuse to embrace AI if it is under the control of only a handful of powerful American companies.[33]

Some employees from Google, OpenAI, and others have been critical of Meta and the broader push towards increasing access to models,[34] saying that allowing fully open access to cutting-edge AI technology is dangerous. Despite these concerns, reports circulated in May 2023 indicated that OpenAI was preparing to release a new open source LLM to the public.[35] Hugging Face also announced the launch of an open source alternative to ChatGPT called HuggingChat,[36] although it remains unclear whether HuggingChat can be used commercially due to licensing issues (the model is based on Meta's LLaMA, which lacks a commercial license).

[28] Dylan Patel and Afzal Ahmad, "Google 'We Have No Moat, and Neither Does OpenAI,'" *SemiAnalysis* (blog), May 4, 2023, https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

[29] Thite Anish (@anish), "Llama's Running at 5 Token / Sec on a Pixel 6 Thanks to @koansin," Twitter, March 14, 2023, https://twitter.com/thiteanish/status/1635678053853536256.

[30] Tloen, Tloen/Alpaca-Lora, "Instruct-Tune Llama on Consumer Hardware," GitHub repository, accessed November 6, 2023, https://github.com/tloen/alpaca-lora.

[31] Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models," Centre for the Governance of AI, September 29, 2023, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

[32] Stability AI, "Advocating for Open Models in AI Oversight: Stability AI's Letter to the United States Senate," press release, May 16, 2023, https://stability.ai/blog/stability-ai-letter-us-senate-ai-oversight.

[33] Cade Metz and Mike Isaac, "In Battle over A.I., Meta Decides to Give Away Its Crown Jewels," *New York Times*, May 18, 2023, https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html.

[34] Irving Wladawsky-Berger, "Are Open AI Models Safe?" The Linux Foundation (blog), June 2, 2023. https://www.linuxfoundation.org/blog/are-open-ai-models-safe.

[35] Ananya Mariam Rajesh, "OpenAI Readies New Open-Source AI Model, the Information Reports," *Reuters*, May 15, 2023, https://www.reuters.com/technology/openai-readies-new-open-source-ai-model-information-2023-05-15/.

[36] Hugging Face, Huggingchat, version 6, accessed November 6, 2023, https://huggingface.co/chat/.

In early July 2023, details about OpenAI's GPT-4 leaked online,[37] indicating that the model had over 10 times more parameters than its predecessor, GPT-3. A week later, Meta and Microsoft jointly released the aforementioned LLaMA 2, the next generation of their open-source LLM, along with a "cookbook"[38] for fine-tuning the model.[39] LLaMA 2 is free to download for both research and commercial use.

In December 2023, more than 50 organizations led by Meta and IBM signed on to the AI Alliance, a "community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation."[40] Alliance members note the need for responsible approaches to tackling risk, and remain committed to working on open foundation models, including highly capable multilingual, multi-modal, and science models, marking a collective push against companies building closed models like OpenAI, Google, and Anthropic.[41]

Public Discourse on Model Access

Since the Google memo leak, public discourse around increased access to AI models has exploded in volume and pushed the dialogue into two primary and historically familiar camps: those in favor of allowing widespread and increased access to cutting-edge models, and those concerned about the potential risks associated with unmediated access. Some make the argument that responsibility lies not with the developers of foundation models, but with app developers and users instead. Governments and policymakers are also leaning into this conversation.

While not specifically referencing access to models and their components, the Cybersecurity and Infrastructure Security Agency (CISA) Director Jen Easterly indicated grave concern about AI risks in a May 2023 speech,[42] and in a Judiciary subcommittee hearing with representatives of leading AI companies that same month, Missouri Senator Josh Hawley asked, "[w]ill we strike that balance between technological innovation and our ethical and moral responsibility

[37] Damir Yalalov, "GPT-4's Leaked Details Shed Light on Its Massive Scale and Impressive Architecture." *Metaverse Post*, July 11, 2023, https://mpost.io/gpt-4s-leaked-details-shed-light-on-its-massive-scale-and-impressive-architecture/.

[38] ScaleAPI, "Fine-Tuning Using LLM Engine," GitHub repository, accessed November 6, 2023, https://github.com/scaleapi/llm-engine/blob/main/docs/examples/finetuning.ipynb.

[39] Meta, "Meta and Microsoft Introduce the Next Generation of Llama," press release, July 18, 2023, https://about.fb.com/news/2023/07/llama-2/.

[40] AI Alliance, "AI Alliance," accessed December 2, 2023, https://thealliance.ai.

[41] Ryan Heath, "Open source AI fights back," *Axios*, December 5, 2023, https://www.axios.com/2023/12/05/open-source-ai-fights-back.

[42] Christian Vasquez, "Top US Cyber Official Warns AI May Be the 'Most Powerful Weapon of Our Time,'" *CyberScoop*, May 5, 2023, https://cyberscoop.com/easterly-warning-weapons-artificial-intelligence-chatgpt/.

to humanity?"[43] While these comments were not focused on the access component exclusively, they highlight growing concern among policymakers and government entities about AI-driven risks. As confirmed through the working group process, these risks directly intersect and interact with increased access.

Concern about potential risks of increased access to AI models has also expanded to the European Union (EU) context, where the AI Act was voted out of committee in early May.[44] While the Act has not yet been adopted, the agreed-upon legislation subjects some proprietary foundation models to restrictions, while granting broad exemptions to open source models..[45]

In June 2023, French President Emmanuel Macron announced new funding for an open "digital commons" for French-made generative AI projects, saying "On croit dans l'open-source"[46] (We believe in open-source) and putting the weight of the French government behind the open access movement. Also in June 2023, Senate Majority Leader Chuck Schumer led a bipartisan coalition in writing a letter to coalesce support around ensuring the safe and ethical development and deployment of AI.[47] The letter stressed the challenges posed by a lack of technical knowledge in efforts to create regulatory and other frameworks around AI, stating "[t]he advances we have seen in Artificial Intelligence (AI) in the last few months have been astounding…[a]s AI transforms our world, the Senate must keep abreast of the extraordinary potential, and risks, AI presents."[48]

In July 2023, the Cyberspace Administration of China (CAC) unveiled a set of AI guidelines that encourage the innovative use of generative AI across industry,[49] urge companies to participate in the formulation of international norms for generative AI, and highlight China's support for the development of "secure and trustworthy" chips, software, data, and computing

[43] Office of Senator Josh Hawley, "Hawley Co-Chairs A.I. Judiciary Subcommittee Hearing, Raises Concerns about Election Integrity and Company Liability," press release, May 16, 2023, https://www.hawley.senate.gov/hawley-co-chairs-ai-judiciary-subcommittee-hearing-raises-concerns-about-election-integrity-and.

[44] Samuel Oba, "Amended EU AI Act Takes Aim at American Open-Source AI Models and API Access," Medium, May 15, 2023, https://medium.com/coinmonks/amended-eu-ai-act-takes-aim-at-american-open-source-ai-models-and-api-access-c515fe47e3d2.

[45] Anthony Faiola, Cat Zakrzewski, and Beatriz Ríos, "E.U. Reaches Deal on Landmark AI Bill, Racing Ahead of U.S.," *The Washington Post*, December 8, 2023, https://www.washingtonpost.com/technology/2023/12/08/ai-act-regulation-eu/#:~:text=The%20legislation%20ultimately%20included%20restrictions,their%20own%20products%20and%20tools.

[46] Mohar Chatterjee and Gian Volpicelli, "France bets big on open-source AI," *Politico*, August 4, 2023, https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/.

[47] Senate Democrats, "Leader Schumer Leads Bipartisan Dear Colleague Letter – with Senators Rounds, Heinrich, and Young – Announcing Three Bipartisan Senators-Only Briefings This Summer, Including First-Ever Classified All-Senators AI Briefing: Senate Democratic Leadership," press release, June 6, 2023, https://www.democrats.senate.gov/newsroom/press-releases/leader-schumer-leads-bipartisan-dear-colleague-letter_with-senators-rounds-heinrich-and-young--announcing-three-bipartisan-senators-only-briefings-this-summer-including-first-ever-classified-all-senators-ai-briefing.

[48] Senate Democrats, "Leader Schumer Leads Bipartisan Dear Colleague Letter."

[49] Ryan McMorrow, Nian Liu and Qianer Liu, "China Updates AI Rulebook," *Financial Times*, July 23, 2023, https://www.ft.com/content/eec1d271-b479-4928-870d-3a6a895cffde#post-2cdb10c6-1196-42c6-80e5-79f64dd2de2f.

power.[50] However, the Chinese government and these guidelines have yet to address the question of open access.

On October 30, 2023, U.S. President Joseph Biden issued Executive Order (E.O.) 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,[51] focusing heavily on so-called "dual use" models—a clear nod to technologies that have military applications and thus are subject to export controls. While the order does not specifically address the open access question, Section 4.6 seeks to understand the implications of models with "widely available weights," indicating a focus on the technical components of models that allow increased access. Among other responsibilities, the E.O. tasks the National Telecommunications and Information Administration (NTIA) with preparing a report to assess the risks and benefits that arise when model weights are open sourced.[52] Further, the E.O. directs the Secretary of Commerce, acting through the Director of the National Institute of Standards and Technology (NIST) to establish the U.S. Artificial Intelligence Safety Institute to lead efforts on AI safety and trust. The Institute will develop guidelines and best practices with the aim of promoting consensus industry standards for developing and deploying safe, secure, and trustworthy AI systems. The Institute's work will no doubt have an impact on the debate around open access and potential risk.[53]

[50] Laura He, "China Takes Major Step in Regulating Generative AI Services like ChatGPT," *CNN*, July 14, 2023, https://www.cnn.com/2023/07/14/tech/china-ai-regulation-intl-hnk/index.html.

[51] Executive Order No. 14100, 88 Federal Register 75191, October 30, 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[52] U.S. Department of Commerce, "Department of Commerce to Undertake Key Responsibilities in Historic Artificial Intelligence Executive Order," press release, October 30, 2023, https://www.commerce.gov/news/press-releases/2023/10/department-commerce-undertake-key-responsibilities-historic-artificial.

[53] U.S. Department of Commerce, "At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety," press release, November 1, 2023, https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial.

Categories of Opportunity and Risk

Continued AI advancements and adoption will deliver significant benefits across the global economy, but also introduce risks that are less well understood and, admittedly, somewhat speculative. While this report is focused on the risks of access to powerful AI foundation models and their components, we also provide a summary of some of the benefits of open access to these capabilities as identified by the working group and our independent research for context and balance.

Identified Opportunities

AI's potential uses and benefits are difficult to fully imagine, let alone represent, but minimally include improved cyber defenses, increased scientific development, expedited economic development, and improved quality of life.

» **Improved cyber defenses:** Concern that malicious actors will leverage AI to carry out a range of cybercriminal activities is at least in part offset by the potential of AI to improve cyber defenses.[54] AI can provide flexible decision making mechanisms and computational intelligence capabilities to defensive systems and increase human capacity to scan for and identify vulnerabilities.[55] It is worth noting that this increased detection capacity will only translate to defensive gains if accompanied by a restructuring of existing vulnerability management and mitigation processes and procedures, to include training the humans that will continue to be required for system oversight and management.

» **Increased scientific development:** AI holds the promise of accelerating progress in scientific discovery, including within the fields of medicine, climate science, and beyond.[56] It could also facilitate collaboration and knowledge sharing among researchers and developers by enabling the collection, processing, and storage of data in distributed

[54] Enn Tyugu, "Artificial Intelligence in Cyber Defense," *2011 3rd International Conference on Cyber Conflict* (2011): 1-11, https://ieeexplore.ieee.org/document/5954703/authors.

[55] Selma Dilek, Hüseyin Çakır, and Mustafa Aydın, "Applications of artificial intelligence techniques to combating cyber crimes: A review," arXiv, January 1, 2015, https://arxiv.org/pdf/1502.03552.pdf.

[56] Yongjun Xu et al., "Artificial Intelligence: A Powerful Paradigm for Scientific Research," *The Innovation* 4, no.2 (2021): 100179, October 28, 2021, https://www.cell.com/article/S2666-6758(21)00104-1/fulltext.

and crowd-sourced environments.[57] Access to these capabilities has and will continue to dramatically expand the ability of an increasing number of developers to push these frontiers.

- **Expedited economic development:** AI holds the potential to alleviate some costs and other burdens associated with economic development, to include those faced by low- and middle-income countries, by facilitating bespoke or low-cost solutions to a range of economic challenges.[58] AI might increase labor productivity,[59] create a new "virtual workforce" capable of solving problems and self-learning, and encourage a diffusion of technological innovation, creating new revenue streams.[60]
- **Improved quality of life:** AI has the potential to improve individual lifestyles, for example through AI personal assistants to save time,[61] personalized medicine to increase health and longevity,[62] complex problem solving capabilities,[63] scalable coordination between humans and AI technologies enabling collaboration across diverse geographies,[64] and individually tailored uses of products and services.[65]

Identified Risks

For the purposes of this report, the risks posed by the development and proliferation of AI technologies include fueling a race to the bottom, malicious use, capability overhang, compliance failure, taking the human out of the loop, and reinforcing bias.

- **Fueling a race to the bottom:** At the time of writing, an all-out race among leading AI labs is underway to develop and promote their respective foundation models for broader adoption by technology stakeholders. As with other foundational technologies–like operating systems and mobile, cloud, and social media platforms–the common expectation

[57] Xiang Zhang et al., "Geospatial sensor web: A cyber-physical infrastructure for geoscience research and application," *Earth-Science Reviews*, no. 185 (2018): 684-703, https://www.sciencedirect.com/science/article/abs/pii/S0012825217305044.

[58] Jason Furman and Robert Seamans, "AI and the Economy," *Innovation Policy and the Economy*, no. 19 (2019): 161-191, https://www.journals.uchicago.edu/doi/full/10.1086/699936.

[59] Martin Neil Baily, Erik Brynjolfsson, and Anton Korinek, "Machines of mind: The case for an AI-powered productivity boom," *Brookings*, May 10, 2023, https://www.brookings.edu/articles/machines-of-mind-the-case-for-an-ai-powered-productivity-boom/.

[60] European Parliament, *Economic Impacts of Artificial Intelligence*, by Marcin Szczepański, PE 637.967, 2019, https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf.

[61] Nil Goksel Canbek and Mehmet Emin Mutlu, "On the Track of Artificial Intelligence: Learning with Intelligent Personal Assistants," *Journal of Human Sciences* 13, no. 1 (2016): 592–601, https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/3549.

[62] Nicholas J. Schork, "Artificial Intelligence and Personalized Medicine," in *Precision Medicine in Cancer Therapy*, edited by Daniel D. Von Hoff and Haiyong Han (Springer, Cham, 2019), 265-283, https://doi.org/10.1007/978-3-030-16391-4_11.

[63] Srecko Joksimovic et al., "Opportunities of Artificial Intelligence for Supporting Complex Problem-Solving: Findings from a Scoping Review," *Computers and Education: Artificial Intelligence*, no. 4 (2023): 100138, https://www.sciencedirect.com/science/article/pii/S2666920X23000176.

[64] Sourav Mondal and Elaine Wong, "Global-Local AI Coordinated Learning over Optical Access Networks for Scalable H2M/R Collaborations," *IEEE Network* 36, no. 2 (2022): 124-130, https://ieeexplore.ieee.org/abstract/document/9785746.

[65] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond, "Generative AI at Work," arXiv, April 23, 2023, https://arxiv.org/abs/2304.11771.

is that the community will likely coalesce around a few leaders in the foundation model adoption race, upon which additive contributions, tailoring, and integrations will be built for the foreseeable future. This dynamic might fuel a race to move products to market as quickly as possible, which could incentivize developer organizations to cut corners in addressing safety, security, and ethics issues.[66] This race to the bottom dynamic is not hypothetical; it has been a persistent concern in the AI space for years, and is now being exacerbated by the breakneck speed and power of innovation.

» **Malicious use:** Unsurprisingly, malicious actors are continuously improving their tools and tradecraft to undermine the safety and security of individuals, groups, or society.[67] Malicious use of AI can result in a range of harms to digital, physical, and political systems, and might include:[68]

  » *Fraud and other crime schemes* enabled by AI-generated social engineering content, particularly when targeting at-risk populations (e.g., children and the elderly);[69]

  » *The undermining of social cohesion and democratic processes* through targeted disinformation campaigns that seek to sow discord or confusion, particularly within the context of elections and political transitions;[70]

  » *Human rights abuses* by expanding the ability of authoritarian states to surveil, constrain, and oppress minorities and dissidents;[71]

  » *Disruption of critical infrastructure* by providing malicious actors with offensive cyber capabilities that outmatch defenses or by introducing cybersecurity vulnerabilities to critical systems;[72] and

  » *State conflict* by contributing to the capabilities of adversarial or revanchist states looking for the means to overcome power and information asymmetries,[73] including through economic, military, intelligence, cyber, cognitive, and/or information operations.

» **Capability overhang:** AI foundation models have exhibited capabilities and aptitudes not envisioned by their developers, commonly referred to as a "capability overhang."

[66] Sebastian Klovig Skelton, "MPs Warned of AI Arms Race to the Bottom," *Computer Weekly*, January 31, 2023, https://www.computerweekly.com/news/365529793/MPs-warned-of-AI-arms-race-to-the-bottom.

[67] Muhammad Mudassar Yamin et al., "Weaponized AI for cyber attacks," *Journal for Information Security and Applications*, no. 57 (March 2021): 102722, https://www.sciencedirect.com/science/article/abs/pii/S2214212620308620.

[68] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv, February 20, 2018, https://arxiv.org/pdf/1802.07228.

[69] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn, "Technical Report: Large Language Models Can Strategically Deceive Their Users When Put Under Pressure," arXiv, November 9, 2023, https://arxiv.org/abs/2311.07590.

[70] Noémi Bontridder and Yves Poullet, "The Role of Artificial Intelligence in Disinformation," *Data & Policy*, November 25, 2021, https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6.

[71] U.K. Government Office for Science, "Future Risks of Frontier AI," October 2023. https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf.

[72] U.K. Government Office for Science, "Future Risks of Frontier AI."

[73] Jinghan Zeng, "Artificial intelligence and China's authoritarian governance," *International Affairs* 96, no. 6 (November 2020): 1441–1459, https://doi.org/10.1093/ia/iiaa172.

Examples include the use of new prompting techniques to enhance model performance, the discovery of new failure modes for AI systems long after their initial release, or more specific cases like simulating a Virtual Machine inside ChatGPT.[74,75] Such unexpected capabilities may also emerge as a result of subsequent fine-tuning. The implications of this risk category remain highly speculative, as it is impossible to evaluate its ramifications without understanding the specific nature of a particular latent, undiscovered capability.

» **Compliance failure:** Verification mechanisms, technical safeguards, and legal guardrails instituted to manage AI risks are only effective when adhered to by developer organizations and users.[76] These controls are increasingly difficult to enforce as the level of access to a model increases, to the point where enforcement may become impossible.

» **Taking the human out of the loop (HOOTL):** Despite its prominence in science fiction, concern that humans might one day lose control of AI systems is legitimate and should be accounted for from the outset. This can include a model's ability to make decisions autonomously and act on them without human verification,[77,78] to strategically deceive users without being instructed to do so,[79] to self-improve or even self-replicate,[80] and to call each others' APIs without human oversight (e.g., UC Berkeley's Gorilla LLM,[81] a fine-tuned LLaMA-based model that can access tools via API calls without a human in the loop). Such concerns were the central theme of IST's February 2023 work on the risks posed by the integration of AI into nuclear command, control, and communications (NC3).

» **Reinforcing bias:** AI models with flawed algorithmic design or trained on biased data have the potential to further entrench existing societal biases and economic inequality.[82,83]

[74] Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, July 6, 2023, https://arxiv.org/abs/2307.03718.

[75] Jonas Degrave, "Building a Virtual Machine inside ChatGPT," *Engraved* (blog), December 3, 2022, https://www.engraved.blog/building-a-virtual-machine-inside/.

[76] Tom Wheeler, "The Three Challenges of AI Regulation," Brookings, June 29, 2023, https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/.

[77] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn, "Technical Report: Large Language Models Can Strategically Deceive Their Users When Put under Pressure," arXiv, November 9, 2023, https://arxiv.org/abs/2311.07590.

[78] Megan Kinniment et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," Alignment Research Center, August 2023, https://evals.alignment.org/language-model-pilot-report/.

[79] Scheurer, Balesni, and Hobbhahn, "Technical Report: Large Language Models Can Strategically Deceive Their Users When Put under Pressure."

[80] "Evaluating the Capabilities and Alignment of Advanced ML Models," Alignment Research Center, accessed November 6, 2023, https://evals.alignment.org/#:~:text="Autonomous%20replication"%20capabilities%20means%20an,seem%20necessary%20for%20autonomous%20replication.

[81] Shishir G. Patil et al., "Gorilla: Large Language Model Connected with Massive APIs," arXiv, accessed November 6, 2023, https://gorilla.cs.berkeley.edu/.

[82] Shahriar Akter et al., "Algorithmic Bias in Data-Driven Innovation in the Age of AI," *International Journal of Information Management*, no. 60 (July 2021): 102387, https://www.sciencedirect.com/science/article/abs/pii/S0268401221000803.

[83] Pauline Kim, "AI and Inequality," *Social Science Research Network*, October 11, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3938578.

Gradient of Access to AI Foundation Models

AI models are composed of a number of components, all of which are required to train and run them.[84] This makes access considerations more complex than a binary "open" or "closed" distinction that may be appropriate in other contexts, like software, and which is often used to describe access in colloquial terms. Developer organizations can choose to share any combination of these components with any subset of users, resulting in a wide range of potential risks and opportunities.[85] The most commonly released model components include:[86]

- M*odel architecture:* Code that specifies the structure and design of an AI model.
- *Model weights:* Variables or numeric values used to specify how an input is transformed into an output.
- *Inference and training code:* Used to implement the trained model.
- *Algorithms:* Used to optimize or fine-tune the model weights during training.

Given the range of components that developers can choose to release publicly or retain internally, a gradient of access approach is most appropriate to describe levels of access.

The gradient we identify draws on a foundational paper by Irene Solaiman that outlines a gradient of generative AI release,[87] along with a number of other critical insights from researchers like Percy Liang and his team at Stanford,[88] and Toby Shevlane,[89] in addition to research from other working group members. Our gradient ranges from "fully closed" proprietary models, in which all aspects and components of the system are inaccessible outside the developer organization, to "fully open" models in which all aspects of the system are accessible and downloadable, including all components. We note that this gradient is necessarily reductive; additional variables that may impact risk, such as release timing

[84] For a more detailed explanation of AI components, see Appendix A of "Open-Sourcing Highly Capable Foundation Models."

[85] Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models," *Centre for the Governance of AI*, September 29, 2023, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

[86] Seger, "Open-Sourcing Highly Capable Foundation Models."

[87] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv, February 5, 2023, https://arxiv.org/abs/2302.04844.

[88] Percy Liang (@percyliang), "Meta's release of OPT is an exciting step towards opening new opportunities for research," Twitter, May 3, 2022, https://twitter.com/percyliang/status/1521627736892010497.

[89] Toby Shevlane, "Structured Access: An Emerging Paradigm for Safe AI Deployment," arXiv, accessed November 6, 2023, https://arxiv.org/ftp/arxiv/papers/2201/2201.05159.pdf.

and distribution constraints other than gating, are not considered.[90] Our resulting gradient identifies seven levels of model access.

LEVEL 0: FULLY CLOSED

Fully closed, proprietary models are those in which all aspects and components of an AI system are inaccessible outside the developer organization, or even restricted to a specific subsection of an organization. Examples include DeepMind's Gopher and Google's Imagen.[91,92]

LEVEL 1: PAPER PUBLICATION

Developer organizations sometimes publish technical papers that provide details about models before or after they are released. Not long ago, the level of transparency in research papers was one of the most contentious topics in this debate, now greatly superseded by the more intense discussions around release of actual models and components themselves. Publishing technical papers on cutting-edge models can contribute to the development of protections against misuse, but can also provide proof that certain model capabilities are possible and reveal general ideas that can be built upon. Further, technical papers may detail a model's training process and code,[93] including configuration settings and telemetry collected during training. While papers differ in technical detail, in some instances they can significantly ease the effort required for actors to replicate and/or fine-tune models. For example, publishing a model's optimal parameters would make a pre-trained AI model more capable (and possibly dangerous),[94] and releasing the code used to clean, label, and load the training data into the model would reduce the burden on actors attempting to reproduce model weights.[95]

[90] Aviv Ovadya and Jess Whittlestone, "Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning," arXiv, July 25, 2019, https://arxiv.org/pdf/1907.11274.pdf.

[91] James Vincent, "Deepmind Tests the Limits of Large AI Language Systems with 280-Billion-Parameter Model," *The Verge*, December 8, 2021, https://www.theverge.com/2021/12/8/22822199/large-language-models-ai-deepmind-scaling-gopher.

[92] "Google AI Foundation Models," Google, accessed November 6, 2023, https://ai.google/discover/foundation-models/.

[93] Solaiman, "The Gradient of Generative AI Release."

[94] George Lawton, "The role of AI parameters in the enterprise," *TechTarget* (2023), helpfully described "parameters" as follows: "One way to understand AI parameters is to picture a cartoon representation of a deep learning neural network with lots of knobs that are wired together. When you present an input to the neural net (say a sentence or an image), these knobs control an enormous number of very simple computations that transform the input into an output via a large number of intermediate steps called layers. When you want to train such a network, you repeatedly present it with an input and the desired output, and you use the difference between the actual output and the desired one as a guide to how to adjust the knobs to make the network do better on this input-output pair in the future….The value of each knob is called a parameter."

[95] Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models," *Centre for the Governance of AI*, September 29, 2023, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models.

The risks and opportunities associated with such publications rest on two primary factors: what exactly the paper contains, and the potential that others could have obtained the same knowledge through other legitimate means, sometimes referred to as counterfactual possession. Ultimately, actors who stand to benefit the most from the publication of technical papers exist in a "Goldilocks zone" of capability: they are sufficiently capable to understand and apply the knowledge outlined in the paper, but not able to independently discover it.[96]

LEVEL 2: QUERY API ACCESS

At the query API access level, AI systems remain hosted on the developer organization's servers, but allow outside users to interact with them directly. Some examples of tasks that can be facilitated via query API access include computer vision (image and object recognition), speech recognition (speech-to-text or vice versa), natural language processing (written question/answer), document parsing, and content generation. OpenAI's ChatGPT-2 is one example of a model with query API access.[97]

Critically, query API access **only allows users to obtain a model's output**, not information about its inputs or how it reached a given output, nor access to model components that may provide additional insight into its "decision making process." While it is possible that users are able to infer some information about how a model reached a given output, for example through smart prompt injections or jailbreaking,[98] acquiring this knowledge requires time and effort, and is not the same as having direct access to these insights.

This level of access typically includes safety measures or filters to prevent users from abusing the system or to prevent the system from producing harmful or undesired outputs, like child sexual abuse material (CSAM). It also allows the developer organization to maintain control over the model's use, allowing for more centralized efforts to screen out harmful inputs, mitigate bias, and manage quality. For example, rate limits imposed in a query API access scenario restrict the number of times a user can access the server within a specific period of time to prevent automated data scraping or malicious use such as denial of service attacks.[99] Developer organizations can also revoke query API access if new vulnerabilities are discovered or unforeseen risks arise.

[96] Toby Shevlane and Allan Dafoe, "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" arXiv, January 9, 2020, https://arxiv.org/abs/2001.00463.

[97] Irene Solaiman et al., "Release Strategies and the Social Impacts of Language Models," arXiv, November 13, 2019, https://arxiv.org/abs/1908.09203.

[98] "Quick Concepts: Jailbreaking," Innodata, accessed October 25, 2023, https://innodata.com/quick-concepts-jailbreaking/#:~:text=Jailbreaking%20is%20a%20form%20of,content%20filters%20would%20otherwise%20block.

[99] "OpenAI Platform," OpenAI, accessed November 6, 2023, https://platform.openai.com/docs/guides/rate-limits?context=tier-free.

LEVEL 3: MODULAR API ACCESS

Modular API access, sometimes also called "advanced API access" or "research API access,"[100] allows users greater insight into models at a similar level of functionality as described in Level 2. This level describes AI systems that remain hosted on the developer organization's servers, but allow outside users to directly interact with them and, importantly, **obtain information about a model's training methodology or inputs**. For example, some developer organizations may choose to grant users access to a range of system components, which may allow users to glean insight into a model's functionality, thereby enabling them to better scrutinize a model's output. Despite this increased access, developer organizations can still revoke modular API access if new vulnerabilities are discovered or unforeseen risks arise.

Modular API access allows research activities to be carried out, but with oversight and controls to prevent the model from being modified or fine-tuned. By hosting models at this level of access on their servers and granting access via APIs, developer organizations are able to monitor and log user interactions, facilitating their ability to trace the origin and evolution of possible harms and prevent users from fine-tuning or replicating models. However, while a researcher with modular API access is unlikely to be able to create a fine-tuned version of the model itself, they may be able to infer sufficient details about the model's components and system information to increase the risk of reverse engineering.

In many cases, models released initially for query API access are later made available to researchers under modular API access by allowing access to specific components or modules of a model over a period of time.[101] This process of gradually increasing access to model components and modules is known as "structured access."[102] OpenAI's GPT-2, for example, was released initially in February 2019 as a 124 million-parameter language model.[103] Over the course of the next nine months, OpenAI released five versions of GPT-2, each time increasing the number of parameters, with the final release in November 2019 featuring a 1.5 billion-parameter version of the model. This allowed time between model releases for researchers to analyze the model's functionality and associated risks and benefits, and ultimately increased OpenAI's confidence in the safety of the models before release. Other developer

[100] Benjamin S. Bucknall and Robert F. Trager, "Structured Access for Third-Party Research on Frontier AI Models: Investigating researchers' model access requirements," AI Governance Initiative and Oxford Martin School, October 27, 2023, https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/.

[101] Toby Shevlane, "Structured Access: An Emerging Paradigm for Safe AI Deployment," arXiv, accessed November 6, 2023, https://arxiv.org/ftp/arxiv/papers/2201/2201.05159.pdf.

[102] Shevlane, "Structured Access."

[103] Irene Solaiman et al., "Release Strategies and the Social Impacts of Language Models," arXiv, November 13, 2019, https://arxiv.org/abs/1908.09203.

organizations have used similar techniques regarding access, such as in the case of GROVER, a model released in June 2019 by the Allen Institute for Artificial Intelligence at the University of Washington, and Hugging Face.[104,105]

→ Downloadable Access

Developer organizations may choose to make a model and its components available for download to user servers with certain restrictions, typically involving the withholding of some system components, such as the model's training datasets or model weights. While a model's size and minimum computing resources will limit who is capable of running it locally (for example, average consumer hardware is unlikely to support large, powerful, downloadable models at the time of writing), capable actors may be able to fine-tune model architecture, remove safety features, and potentially repurpose models for malicious use. It is worth noting that labs are increasingly releasing small but powerful models; in this context, model distillation,[106] or the process of transferring the knowledge from a large model to a smaller model that can be practically deployed under real-world resource constraints, would mitigate the difficulty of running models locally on consumer-grade hardware. Although downloadability may present opportunities for model customization, once models are downloaded into a separate environment, developer organizations can lose control over the ways in which users fine-tune and deploy their models.

LEVEL 4: GATED DOWNLOADABLE ACCESS

Developer organizations offering AI models and components for download will often require users to apply for downloadable access or otherwise register for an account, which can be free or paid. This process is known as "gating" access, and is similar to common practices in news or other content websites where a user might be asked to create a free account. Depending on the stringency of the gating requirements, the developer organization may have some insight into the identity of users who are downloading their models, require users to acknowledge a terms of use agreement, and be able to revoke privileges depending on terms of use.

While better than having no controls whatsoever, gating's ability to prevent undesired outcomes (e.g., unauthorized release or malicious use of a model and its components) is likely limited, not least due to the possibility that gated models might leak online once

[104] Rowan Zellers et al., "Defending against Neural Fake News," arXiv, December 11, 2020, https://arxiv.org/abs/1905.12616.

[105] Clément Delangue, "Ethical Analysis of the Open Sourcing of a State-of-the-Art Conversational AI," Medium, May 9, 2019, https://medium.com/huggingface/ethical-analysis-of-the-open-sourcing-of-a-state-of-the-art-conversational-ai-852113c324b2.

[106] Sundeep Teki, "Knowledge Distillation: Principles, Algorithms, Applications," Neptune.ai (blog), September 29, 2023, https://neptune.ai/blog/knowledge-distillation.

they have been downloaded.[107] This paper will treat gated and non-gated downloadable access separately, as their risk profiles are marginally different. One example of a gated downloadable model is Meta's OPT-175B.[108]

LEVEL 5: NON-GATED DOWNLOADABLE ACCESS

In a non-gated access scenario, a model and certain components released by its developer organization can be downloaded by anyone without subsequent tracking, payment, or terms of use. While a developer organization can choose to withhold some model components and track downloads, they have limited insight into the identity of users downloading the model or the ways in which they are using it.

LEVEL 6: FULLY OPEN ACCESS

In the case of fully open models, all aspects of the system, including all the base model and all components like model weights are accessible and freely downloadable by the public.[109] **Once models are made fully open, it is impossible for developer organizations to walk back a model's release**. For example, BigScience's BLOOM and Mistral's 7B are fully open.[110,111]

[107] James Vincent, "Meta's Powerful AI Language Model Has Leaked Online - What Happens Now?" *The Verge*, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.

[108] "Democratizing Access to Large-Scale Language Models with OPT-175B," Meta (blog), May 3, 2022, https://ai.meta.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/.

[109] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv, February 5, 2023, https://arxiv.org/abs/2302.04844.

[110] Teven Le Scao et al., "Bloom: A 176B-Parameter Open-Access Multilingual Language Model," working paper of Hugging Science, arXiv, June 27, 2023, https://arxiv.org/abs/2211.05100.

[111] Mistral AI, "Mistral 7Bm," September 17, 2023, https://mistral.ai/news/announcing-mistral-7b/.

Matrix: Assessing AI Foundation Model Risk Along a Gradient of Access

With a brief history of the open access movement, a clear understanding of AI's opportunities and risks, and a gradient of access to AI foundation models that exists in practice today, the remainder of this report aims to explain our assessment of how each level of access impacts risk across the six identified categories. While we outline both opportunities and risks in the preceding section, the aim of this project is to reduce harm; as a result, the matrix exclusively focuses on the potential risks of openness. We will revisit potential benefits in the report's conclusion.

Reading the Matrix

The X axis of the matrix represents the gradient of access to AI foundation models and their components from fully closed to fully open as described in the above Gradient of Access to AI Foundation Models section.

The Y axis of the matrix represents the categories of risk as described in the Categories of Opportunity and Risk section. In the Risk Breakdown by Category section that follows the Matrix overview, the Y axis represents the level of risk, from low risk at the bottom to high risk at the top.

Each cell in the matrix captures the effect of a specific level of access on a given category of risk. The colors are largely intuitive:

- 🟦 Relatively low risk
- 🟩 Low-medium risk
- 🟨 Medium risk
- 🟧 Medium-high risk
- 🟥 Relatively high risk

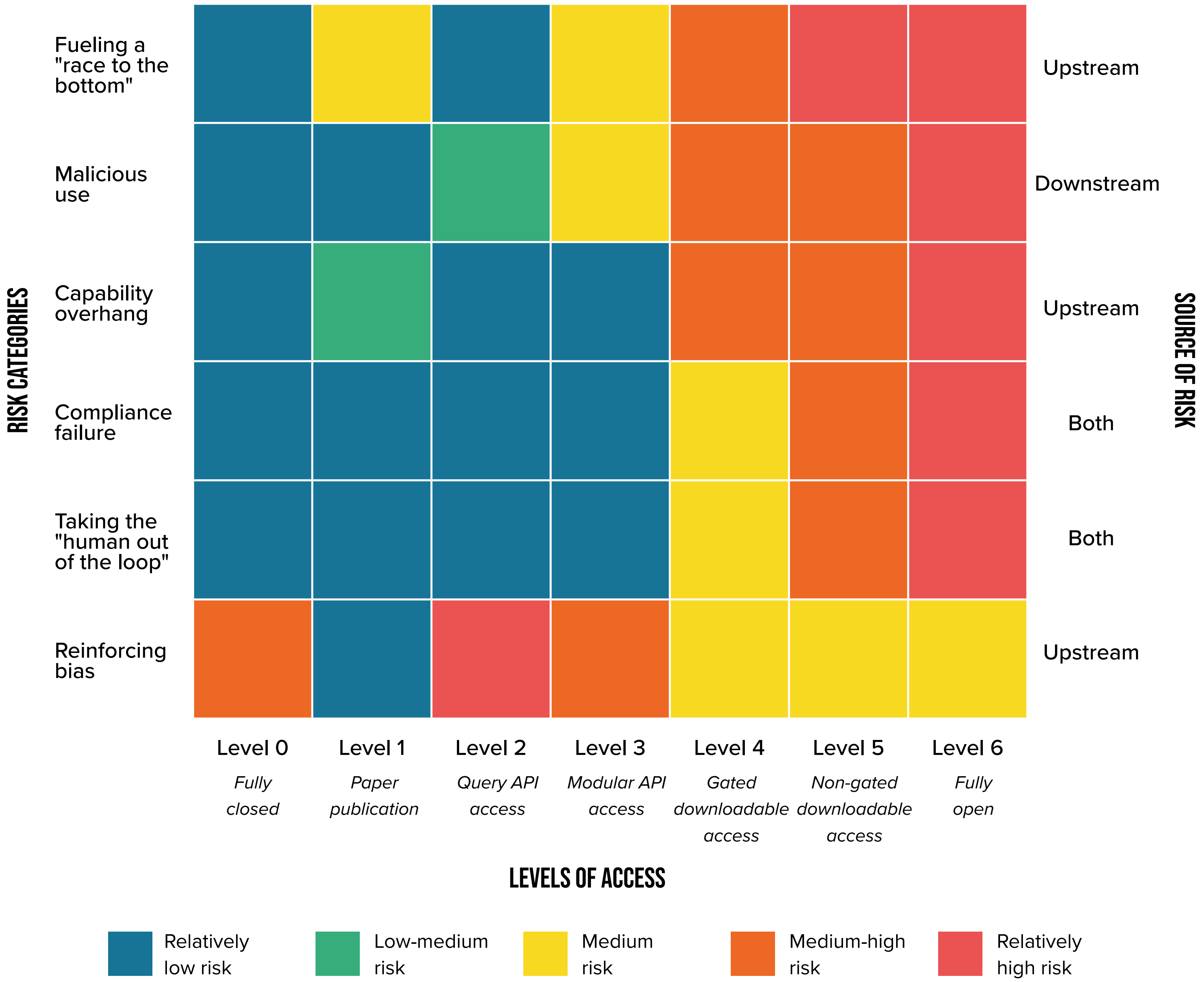
Source of Risk

An additional consideration that surfaced during this project involves the origin of a particular risk. More specifically, source of risk refers to whether the risk is inherent to a model–and therefore a byproduct of the developer organization's design, training, or tuning choices–or driven by a user's (or third party's) interaction with a model at the given level of access. We adopted the terms "upstream" risk for the former and "downstream" risk for the latter, and we use the term **"referent object" to describe the driver of a given risk** (e.g., an AI model is the referent object when a risk is "upstream" and stems from the model itself; a user is the referent object when the risk is "downstream" and is driven by user interaction with a given model).

The source variable is important for two reasons. First, it helps to identify the referent object of the risk in question. **This consideration will become highly relevant when designing technical and policy solutions to mitigate the most extreme risks posed by increased access to cutting-edge AI models.**

Second, this variable impacts our determination of risk at each level of access based on our understanding of a given referent object's available resources (e.g., time, money, compute power, etc.) and technical expertise. In general, leading AI labs have greater resources and technical expertise than users, even in cases where "users" are malicious groups intent on leveraging models to cause harm. As a result, accumulation of upstream risk is generally **the result of inaction or insufficient action** by developer organizations (e.g., not enforcing safety mechanisms), while downstream risk is generally **the result of action** taken by users (e.g., fine-tuning models for malicious purposes). We also note that, in cases where unknowns are present, risk is amplified.

Matrix: Gradient of Access to AI Foundation Models and Associated Risks



RISK CATEGORIES	Level 0 *Fully closed*	Level 1 *Paper publication*	Level 2 *Query API access*	Level 3 *Modular API access*	Level 4 *Gated downloadable access*	Level 5 *Non-gated downloadable access*	Level 6 *Fully open*	SOURCE OF RISK
Fueling a "race to the bottom"	🟦	🟨	🟦	🟨	🟧	🟥	🟥	Upstream
Malicious use	🟦	🟦	🟩	🟨	🟧	🟧	🟥	Downstream
Capability overhang	🟦	🟩	🟦	🟦	🟧	🟧	🟥	Upstream
Compliance failure	🟦	🟦	🟦	🟦	🟨	🟧	🟥	Both
Taking the "human out of the loop"	🟦	🟦	🟦	🟦	🟨	🟧	🟥	Both
Reinforcing bias	🟧	🟦	🟥	🟧	🟨	🟨	🟨	Upstream

LEVELS OF ACCESS

🟦 Relatively low risk

🟩 Low-medium risk

🟨 Medium risk

🟧 Medium-high risk

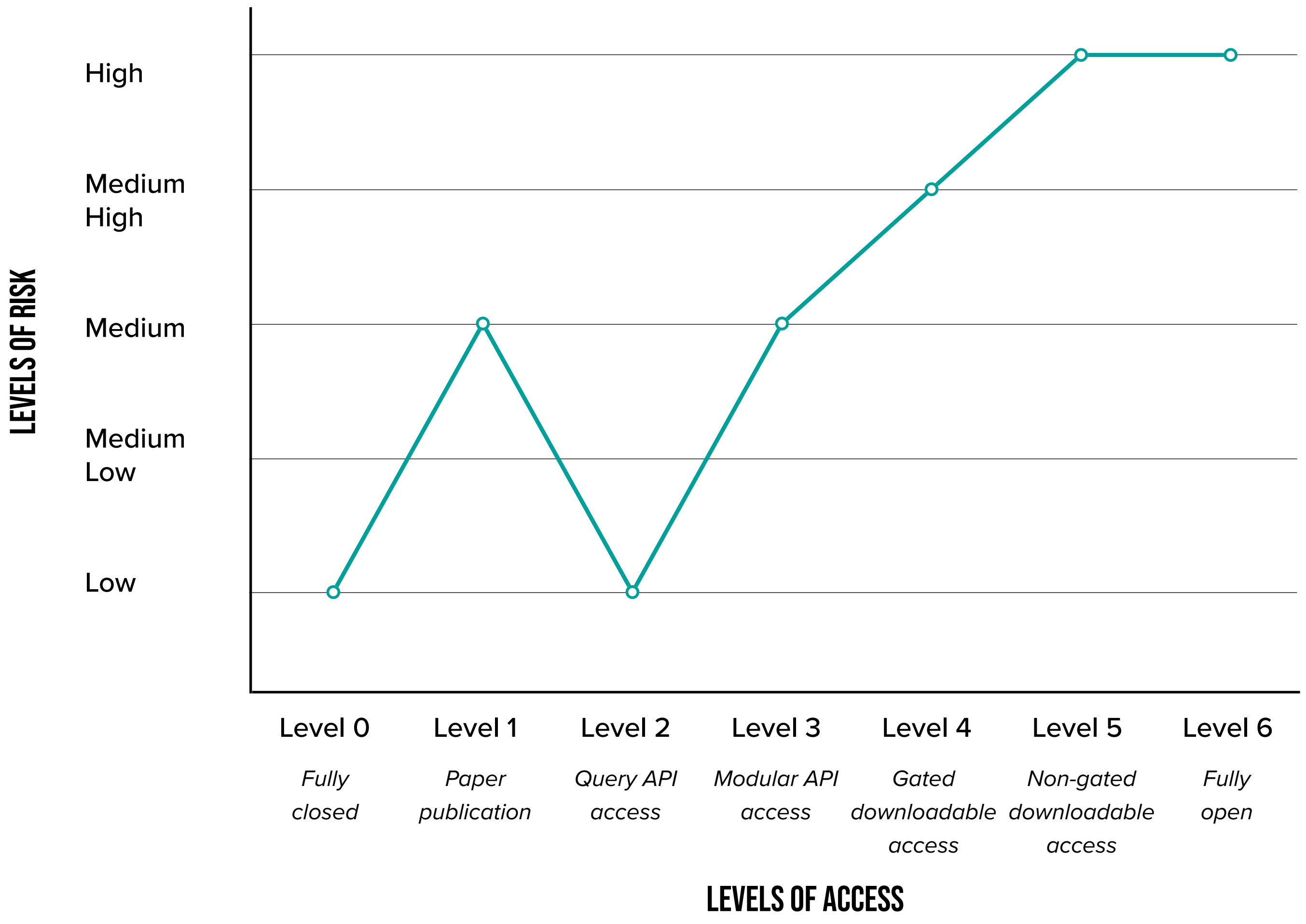
🟥 Relatively high risk

Risk Breakdown by Category

Fueling a race to the bottom

Referent object: AI labs	**Source of risk:** upstream

RISK OF FUELING A RACE TO THE BOTTOM ALONG A GRADIENT OF ACCESS



As outlined in the History of Access to AI Foundation Models section of this report, discussions about access to these models have fractured along two major viewpoints: those in favor of democratizing access to cutting-edge models and those concerned about the associated risks. Developer organizations similarly tend to fall in one category or the other, with some companies aligning their business models with an open access approach and others pursuing a traditional closed/proprietary approach. One argument against allowing widespread access is the risk that such access might fuel a "race to the bottom" in which developer organizations are increasingly incentivized to cut corners in model development in order to quickly release new models to stay competitive. This category of risk describes that concern.

We note that race to the bottom dynamics are highly complex and can be difficult to predict. In the context of this paper, we assume a "winner takes all" dynamic, where a model winning broad adoption will gain market share–and thus future profit opportunities–at the expense of competitors.

At the **fully closed** (Level 0) and **query API access** (Level 2) levels, the risk of fueling a race to the bottom is *low* because only limited information about model components and performance is publicly available, and it is resource intensive for developer organizations to reverse-engineer this information at these levels of access.

Depending on the content of a **published paper** (Level 1), developer organizations may be able to derive significant information about a model's capabilities and identify areas to cut corners in order to achieve a more competitive position vis-à-vis other developer organizations, resulting in a *medium* level of risk.

Modular API access (Level 3) can enable developer organizations to obtain a deeper level of knowledge about models and their outputs than query API access, increasing the possibility that they are able to derive non-released components and other system information. This information might then be used to cut corners in the development of new models to maintain competitive advantage, resulting in *medium* risk of a race to the bottom.

The risk of a race to the bottom increases when models are downloadable because the baseline for technological development essentially rises to the most powerful downloadable model of the day. The "gating" component remains relevant because it may enable some developer organizations to withhold downloadable models from their competitors. As a result, the risk of a race to the bottom is *medium-high* for **gated downloadable** models (Level 4), and *high* for **non-gated downloadable** models (Level 5).

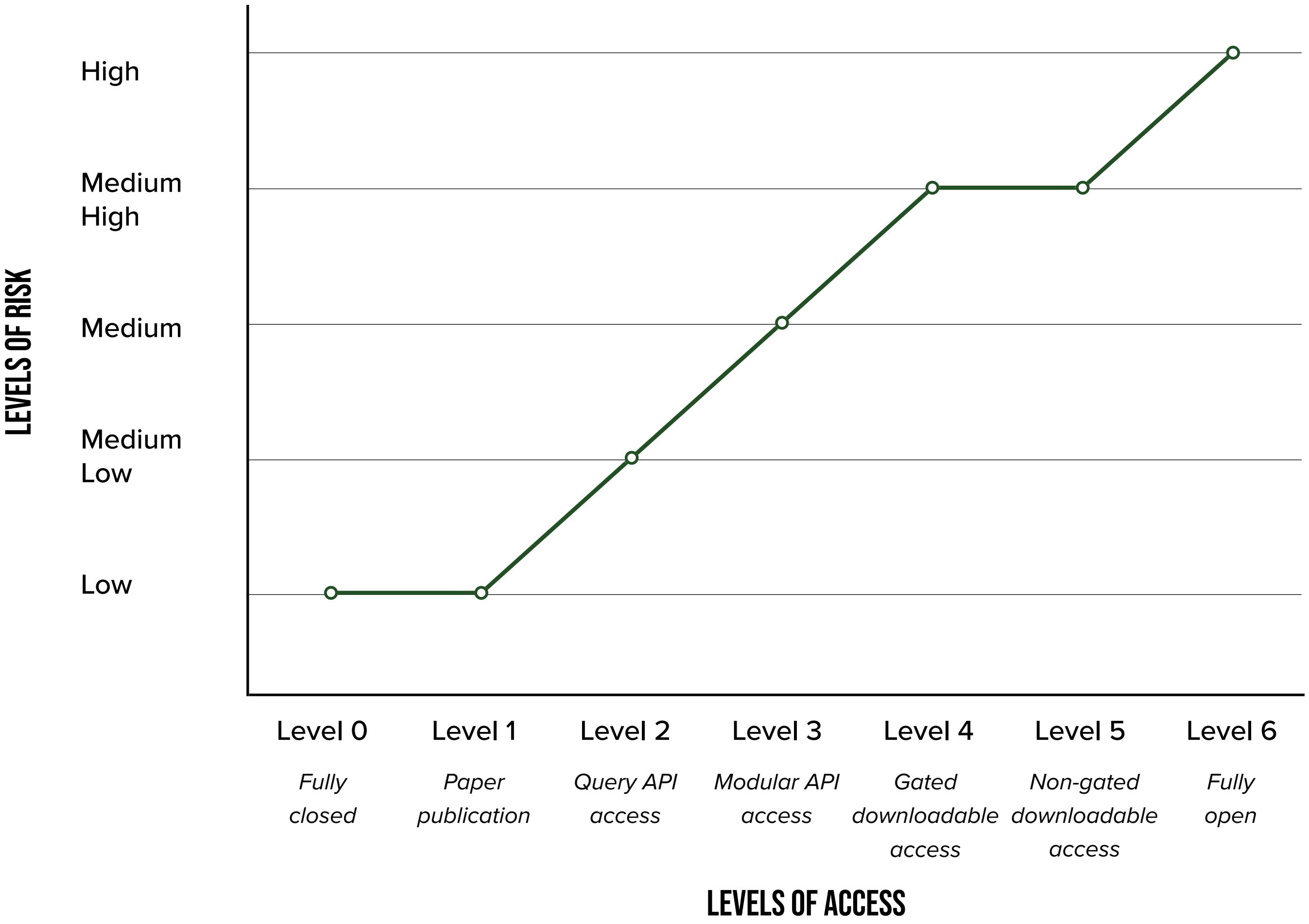
Fully open models (Level 6) present a *high* risk for a race to the bottom when assuming a "winner takes all" dynamic, because they allow all developer organizations to access the most cutting-edge fully open models of the day. Maintaining competitiveness in an environment with many fully open models may therefore require leading labs to cut corners in model development.

Malicious use

Referent object: users

Source of risk: downstream

RISK OF MALICIOUS USE ALONG A GRADIENT OF ACCESS



Malicious use occurs when actors leverage AI to undermine the safety and security of individuals, groups, or society, and can result in a range of harms to digital, physical, and political systems. As the matrix illustrates, as access to cutting-edge AI models increases, so too does the risk of malicious use. At the **fully closed** (Level 0) and **paper publication** (Level 1) access levels, the risk of malicious use is *low* due to the extensive resources and technical expertise required to develop, tweak, or reverse-engineer a model.

The risk of malicious use increases slightly with **query API access** (Level 2) to *low-medium*. A user with this level of access might attempt to bypass safeguards that would ordinarily prevent the model from generating harmful content—like bomb-making instructions, fraudulent content, or sexual abuse material–through creative prompting, a process referred to as jailbreaking.[112] Risk increases slightly to *medium* with **modular API access** (Level 3)**,** which grants deeper insight into models and their capabilities and may reduce both the resource and technical burdens on actors to develop, tweak, or reverse-engineer models for malicious purposes.

The risk of malicious use increases again to *medium-high* when models are **downloadable** (Levels 4 and 5). This is because the resources and technical expertise required to develop, tweak, or reverse-engineer a model decreases significantly when an actor is in possession of the model's architecture and available components. While **gating** (Level 4) can provide a level of traceability for developer organizations and accountability for those downloading a model and its components, preventing malicious use is ultimately more limited for the reasons discussed in the Gradient of Access to AI Foundation Models section.

Finally, the risk that **fully open** (Level 6) models might facilitate malicious use is *high*, as malicious actors are free to develop, tweak, and fine-tune models for malicious purposes to the extent they have the resources to do so. A recent report, for example, describes how researchers effectively removed the safety fine-tuning from Llama 2-Chat 13B with less than $200, thereby demonstrating that when model weights are released publicly, safety fine-tuning is ineffective at preventing misuse.[113]

→ AI's impact on the cyber offense/defense balance

While AI can, and undoubtedly will, be employed by bad actors to advance their offensive aims, it must be noted that AI can also be employed on the defensive side to find and reduce vulnerabilities and support network defense operations. This report does not explore the potential impact that AI could exert on shifting the offense/defense balance or how the openness gradient bears on it. IST will focus on this topic in a separate research project.

[112] "Quick Concepts: Jailbreaking," Innodata, accessed October 25, 2023, https://innodata.com/quick-concepts-jailbreaking/#:~:text=Jailbreaking%20is%20a%20form%20of,content%20filters%20would%20otherwise%20block.

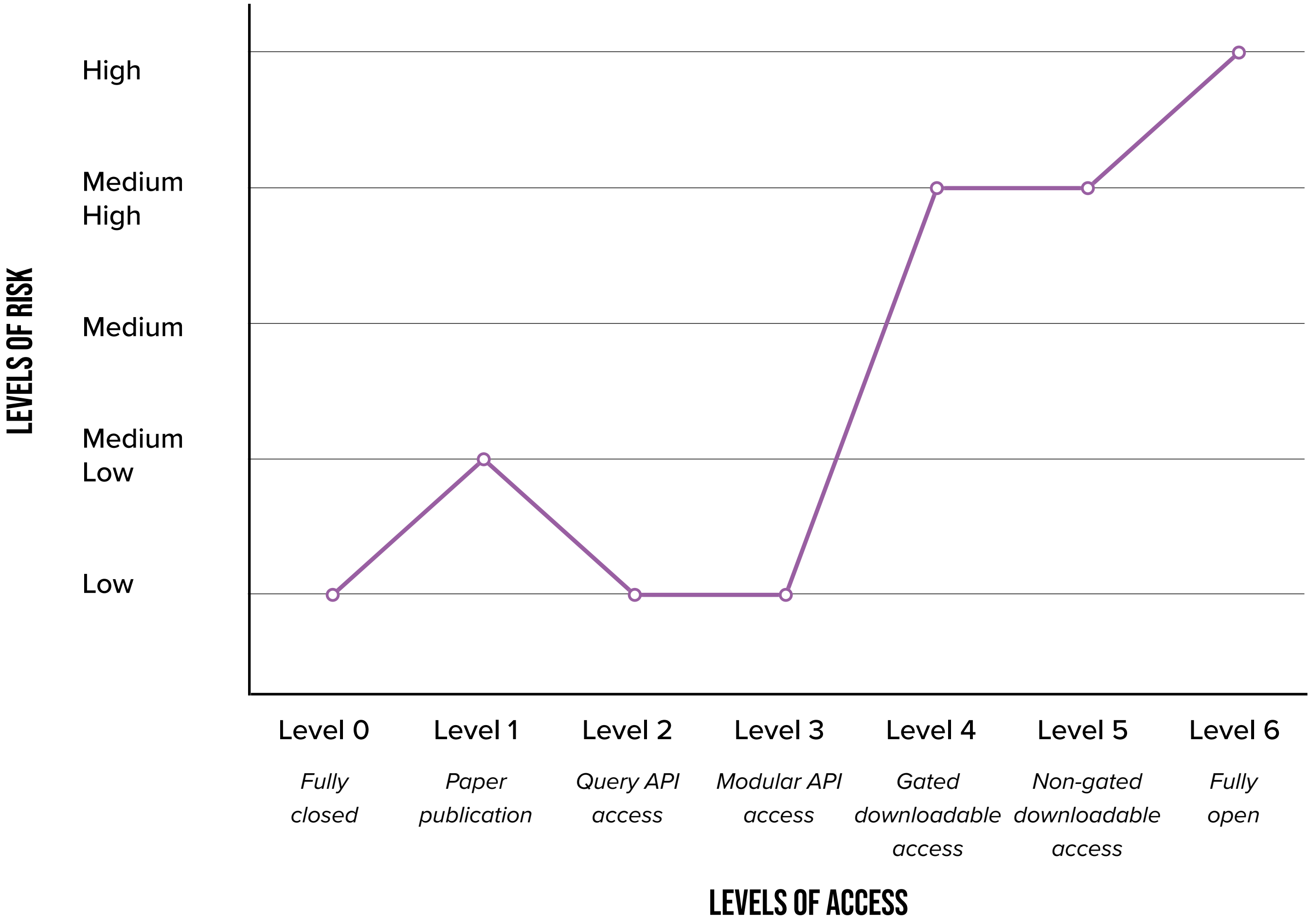
[113] Pranav Gade et al., "BadLlama: Cheaply removing safety fine-tuning from Llama 2-Chat 13B," arXiv, October 31, 2023, https://arxiv.org/abs/2311.00117.

Capability overhang

Referent object: AI models and developer organizations

Source of risk: upstream

RISK OF CAPABILITY OVERHANG ALONG A GRADIENT OF ACCESS



Capability overhang refers to capabilities and aptitudes exhibited by AI foundation models that were not envisioned by their developers. The implications of this risk category remain speculative, as it is impossible to evaluate its ramifications without understanding the specific nature of a particular latent capability until it is discovered. This analysis therefore relies on the limited available research describing when and how capability overhangs arise and on the consensus of our working group.

The relationship between the risks fueled by an AI model's capability overhang and the level of access to the model hinges on three factors: (1) potential for an unexpected capability to manifest in the first place; (2) potential for it to drive other risks and harms (e.g., malicious use or taking the human out of the loop); and (3) potential that the developer organization will identify and mitigate it prior to the onset of undesired outcomes.

The capability overhang risk associated with **fully closed** (Level 0) models is *low*, as such models are proprietary and unexpected capabilities can be identified and managed by the developer organization. In fact, a model's original developers might welcome discovery of such capabilities in the process of testing the model's limits, which might inspire new use cases and other breakthroughs. However, our working group collectively concluded that most capability overhang discoveries are made by third parties examining or interacting with models that have been released externally (whether intentionally or not).

Publishing a technical **paper** (Level 1) on a model slightly increases the capability overhang risk to *low-medium*. While outside researchers are unlikely to detect capability overhang as the result of a paper publication, developer organizations that publish papers have no control over the way the contained information is used.

By contrast, offering users **query and modular API access** (Levels 2 and 3) to a model offers limited potential for capability overhang to be discovered outside of a developer organization's oversight or for this identification to cause negative effects. As a result, the risk dips back to *low*. If developer organizations implement and maintain safety mechanisms (e.g., input format constraints) and track the use of their models, it is unlikely that instances of capability overhang will go undetected. Further, by only offering access through an API, a developer organization maintains control of its model and retains the ability to revoke access to an individual user or more broadly.

The capability overhang risk increases significantly to *medium-high* when models are **downloadable** (Levels 4 and 5). When a model is downloadable, users are better able to tweak, fine-tune, and crowdsource model development, thereby substantially increasing the likelihood that unintended model capabilities may arise. While **gating** (Level 4) can offer a level of traceability and control, it is not necessarily an effective mechanism for technical oversight. It is possible that downloadable access will enable users to identify capability overhangs, and that users may report these instances to developer organizations to be understood and addressed. However, in cases where capability overhang risk exists, developer organizations may not have the ability to independently identify and/or mitigate unintended capabilities as models are downloaded and run locally.

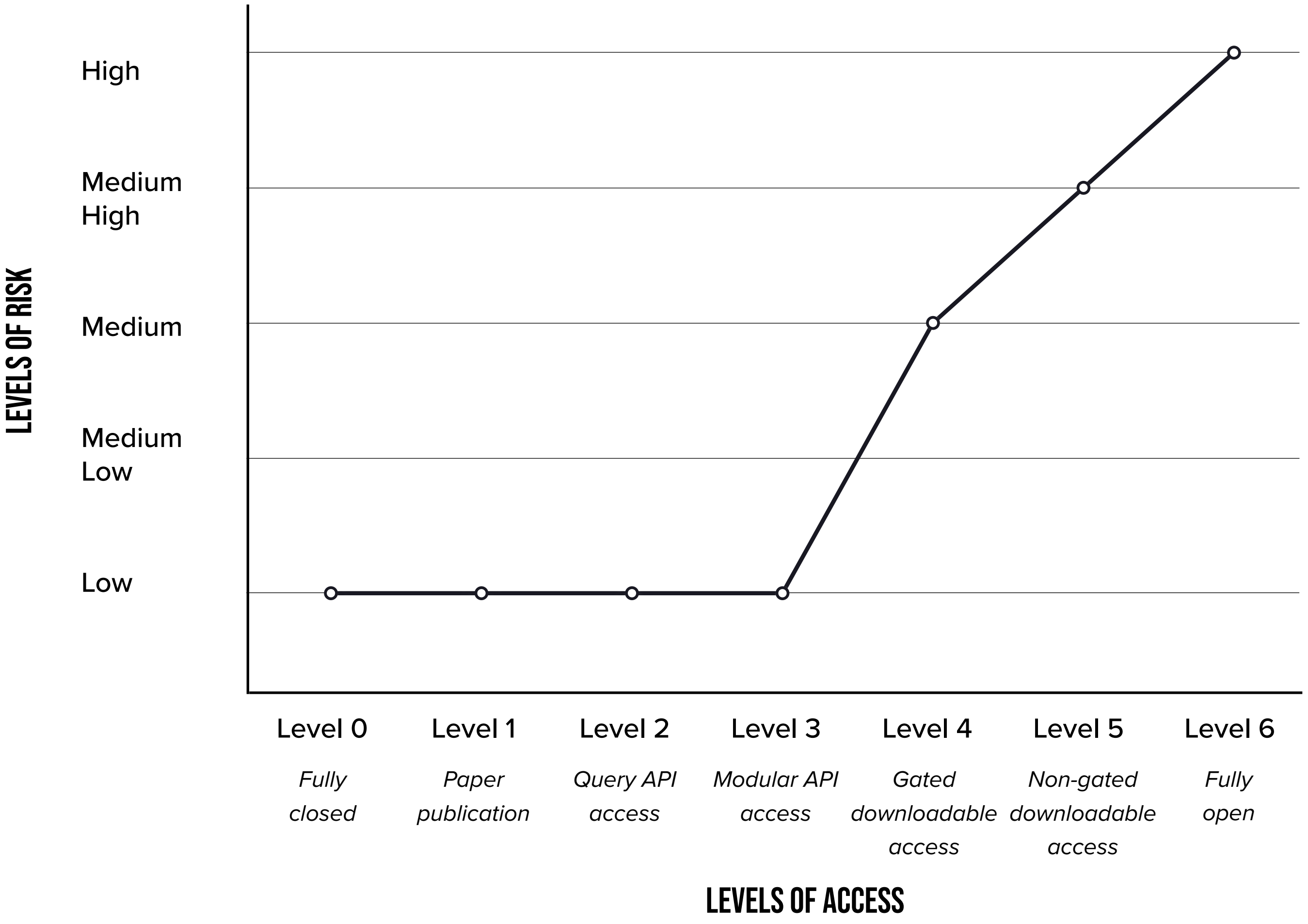
Finally, **fully open** (Level 6) models exhibit *high* capability overhang risk. As more users are able to interact with and tweak models, capability overhang is increasingly likely to occur,[114] and developer organizations are not able to revoke access to these fully open models.

Compliance failure

Referent object: AI models, developer organizations, and users

Source of risk: both upstream and downstream

RISK OF COMPLIANCE FAILURE ALONG A GRADIENT OF ACCESS



Technical, legal, and policy mechanisms established to enforce a model's terms-of-use and prevent its abuse, known as compliance mechanisms, are increasingly diluted as control over a model shifts from the developer organization to users along the gradient of access described in this report. **Fully closed** models, **paper publications**, and **query and modular**

[114] Abhishek Gupta et al., "What ChatGPT Reveals About the Urgent Need for Responsible AI," BCG Henderson Institute, January 19, 2023, https://bcghendersoninstitute.com/what-chatgpt-reveals-about-the-urgent-need-for-responsible-ai/#:~:text=Capability%20overhang%20is%20explored%20by,to%20prompt%20it%20to%20do.

API access (Levels 0, 1, 2, and 3) all pose a *low* risk of compliance failure, as developer organizations carry out “top-down” enforcement, bearing the responsibility for compliance.

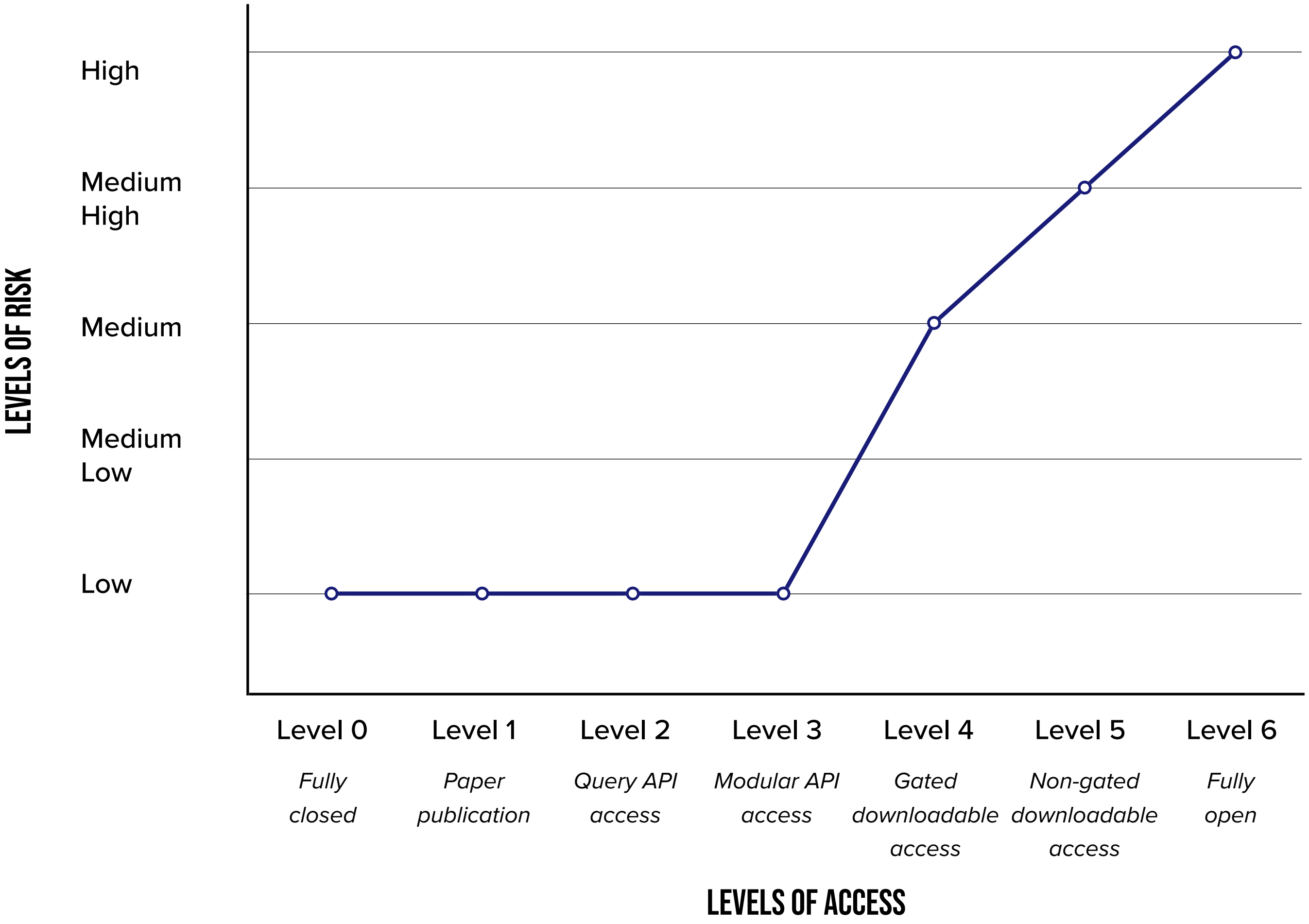
Risk of compliance failure increases steadily to *medium* when models are **downloadable** (Levels 4 and 5), as models are hosted on users’ servers and are beyond the reach of the developer organization. While gating can serve as a mechanism to ensure compliance by restricting model access to self-identified or vetted users (depending on the approach’s details), developer organizations are not truly able to monitor how these models are subsequently employed. When models are available for download without gating (**non-gated,** Level 5), compliance risk slightly increases to *medium-high* by further diluting compliance controls.

Fully open (Level 6) models exhibit a *high risk* of compliance failure, as users have access to all system components and are able to adopt, tweak, and proliferate models beyond the jurisdiction of any enforcement authority.

Taking the human out of the loop

Referent object: AI models and users	**Source of risk:** both upstream and downstream

RISK OF TAKING THE HUMAN OUT OF THE LOOP ALONG A GRADIENT OF ACCESS



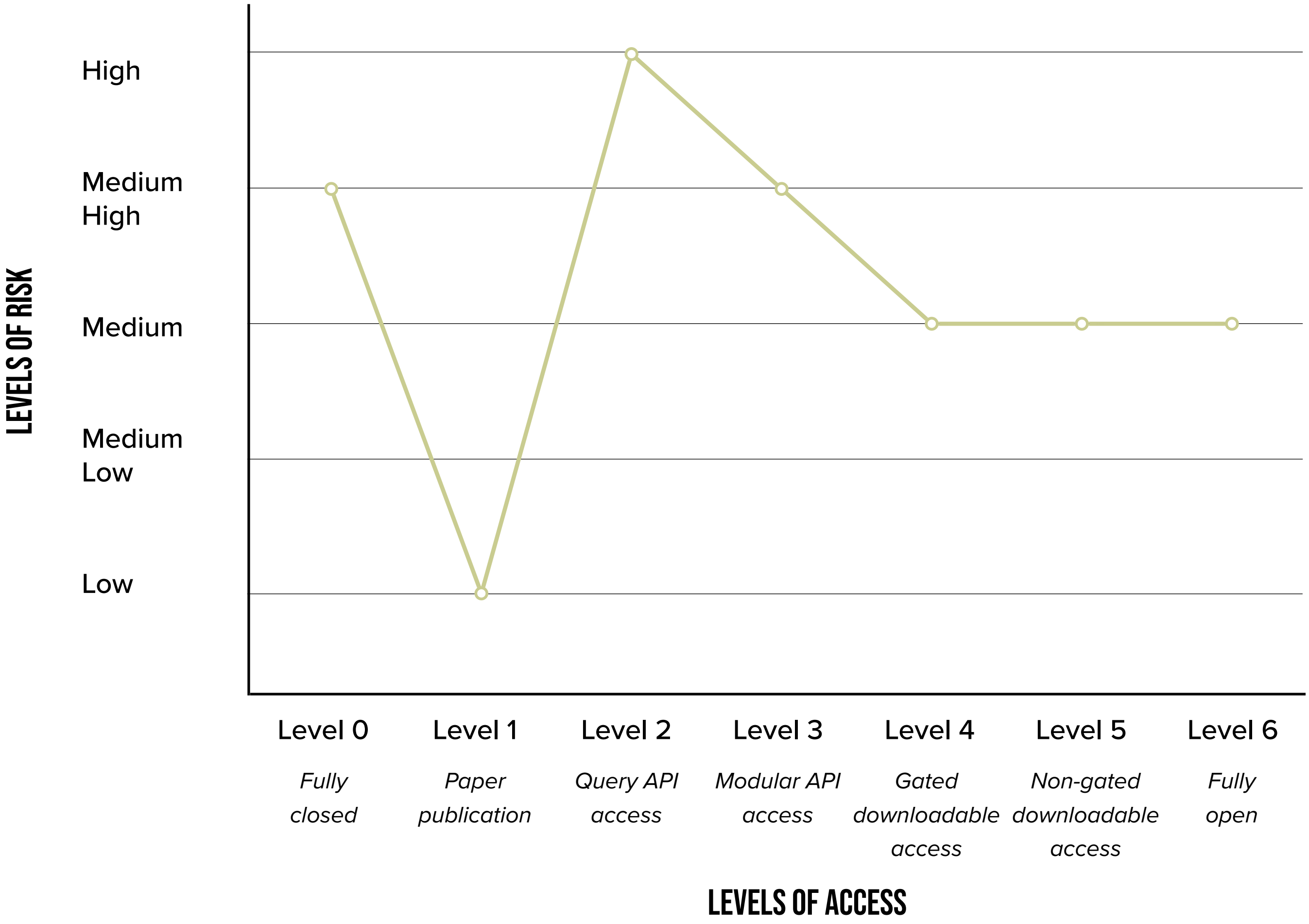
The risk of an AI model making and acting on decisions without human verification or self-improving/self-replicating (including by calling other models via API) without human oversight—thus taking the human out of the loop—is primarily driven by the developer's intended use and corresponding design choices. However, as access to models increases, so too does the opportunity that an actor might download and fine-tune a model in a way that drives this risk. A responsible developer will incorporate safeguards to reduce the likelihood of this possibility, and thus the extent to which they maintain control of the technology will ameliorate the potential of malicious or irresponsible actors altering or repurposing it in a way that allows for such outcomes.

Accordingly, HOOTL risk is *low* for **fully closed**, **paper publication**, and **query and modular API access** (Levels 0, 1, 2, and 3) and increases steadily when models are **downloadable** (Levels 4 and 5) to *medium* when gating is implemented and *medium-high* when ungated. **Fully open** models (Level 6) are at *high* risk of contributing to HOOTL outcomes. We note these conclusions are not based on literature review, because no known study has identified HOOTL outcomes, but instead represent the working group's best estimate based on its collective experience and judgment.

Reinforcing bias

Referent object: AI models and labs	**Source of risk:** upstream

RISK OF REINFORCING BIAS ALONG A GRADIENT OF ACCESS



In the most general sense, as access to a model increases, so too does the potential that third parties might uncover bias resulting from its training data or algorithmic design, thereby creating an opportunity to mitigate the risk. However, even though a developer organization might correct bias in a subsequent model release, there is no assurance that all use of the biased model will cease. Because of this dynamic, the relationship between the risk of reinforcing bias and increased access to models fluctuates.

Fully closed (Level 0) models exhibit a *medium-high* risk of reinforcing bias, as they may contain unknown or undisclosed biases that are propagated to external products or services built upon them, and thereafter to the user base. Transparency is the best innoculation against bias, but is fully lacking in this case by definition.

The act of publishing a technical **paper** (Level 1) on a model tends to lower the risk that it will reinforce bias, depending on the extent to which the paper provides transparency into this risk's potential drivers. Technical papers present a *low* risk of reinforcing bias, unless in the unlikely case that a paper obfuscates or misleads on aspects of a model germane to this risk.

Query API access (Level 2) poses a *high* risk of reinforcing bias. Any risks inherent to models themselves are distributed to the model's user base, the size of which might be quite large given this mode of access (e.g., ChatGPT reportedly has millions of users).[115] The risk that query API outputs might reinforce bias is contingent on the ways in which the developer organization collected data, trained the model, and designed restrictions around inputs and outputs, but the associated risks are distributed to the model's user base.

Modular API access (Level 3) slightly reduces the risk of reinforcing bias to *medium-high* by providing researchers with deeper insight into the models they are interacting with and the outputs they are observing. In some cases, this may allow researchers to identify model bias and alert the developer organization. Additionally, the user base for modular API access is smaller and more sophisticated than that associated with query API access, and less prone to blindly acting on potentially biased output.

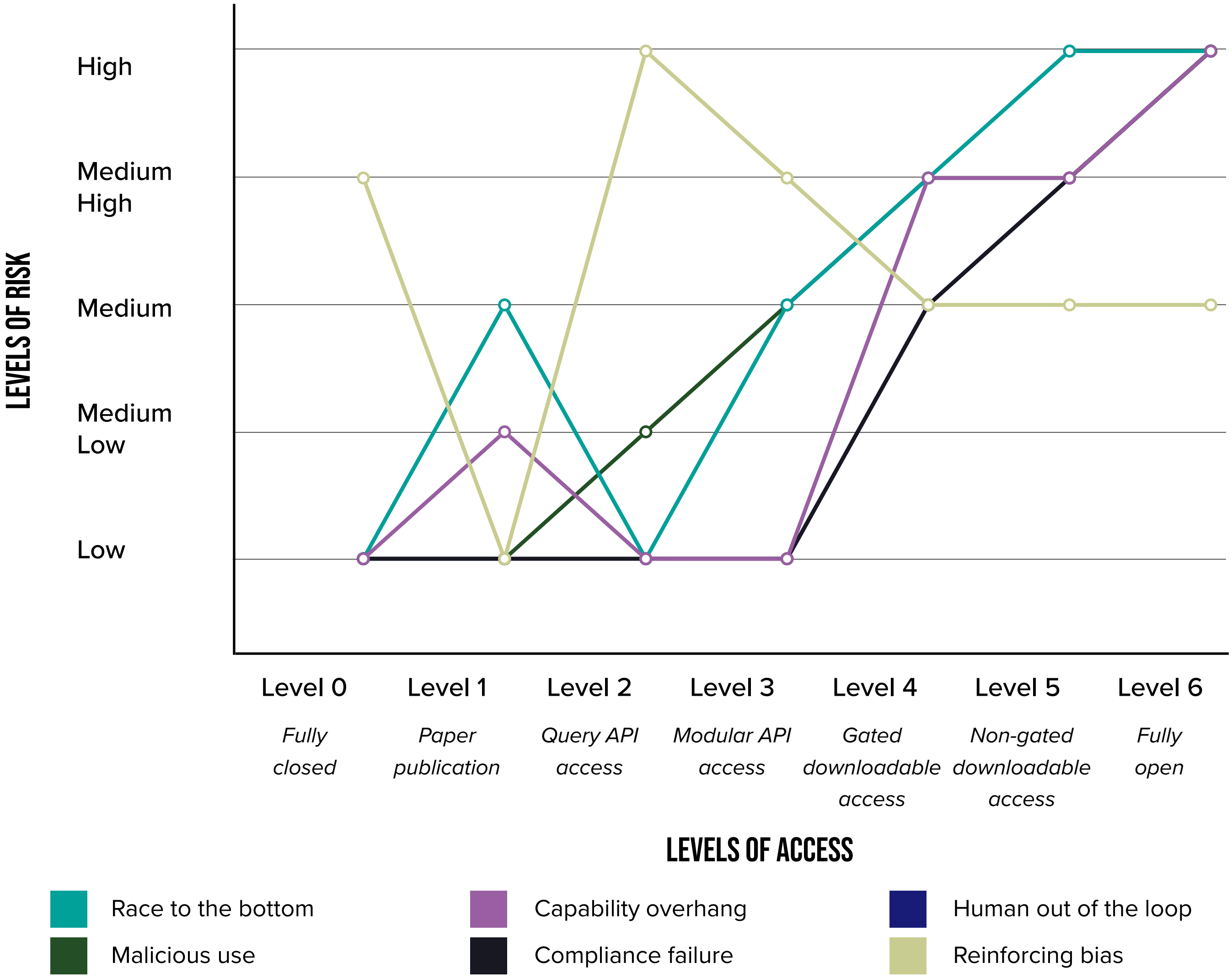
Downloadable (Levels 4 and 5) models present a *medium* risk of reinforcing bias. The number of users interacting with downloadable models is significantly lower than those interacting with models via API because of the resources necessary to run models locally. Further, downloadability provides significantly greater insight into model components that may impact bias. However, even if developer organizations are alerted to unknown biases in their models and choose to update and re-release them to their user base, users may not download the updated model.

[115] Krystal Hu, "ChatGPT sets record for fastest-growing user base - analyst note," *Reuters*, February 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Finally, **fully open** (Level 6) models provide the highest level of transparency, which might enable researchers and developers to identify potential sources of bias in the model itself.[116] However, even if developer organizations are alerted to unknown biases in their models and choose to update and re-release them, at the fully open level these organizations will have no way to identify users operating the original, biased model, and cannot take steps to encourage them to download the updated model. This reality largely offsets any benefits of increased transparency associated with allowing fully open access, resulting in a *medium* risk of reinforcing bias.

Conclusion

OVERVIEW: IDENTIFIED RISKS ALONG A GRADIENT OF ACCESS



[116] Ala Shaabana, "3 Reasons AI Should Be Open Source," *Built In*, November 1, 2023, https://builtin.com/artificial-intelligence/ai-should-be-open-source#:~:text=Open%20Source%20AI%20Can%20Reduce,transparency%2C%20audits%20and%20community%20involvement.

This effort set out to clearly delineate in precise terms where increased access to foundation models and their components increases the risk of harm. The working group discussions and related research pointed to the need for risks to be assessed against a gradient of access, which allows for a more precise identification of where risk arises. In the most general terms, the resulting risk matrix indicates that as access to AI foundation models increases, there is increased potential for harm. More specifically, each category of risk responds somewhat differently to increased levels of access. The risk of malicious use, compliance failure, taking the human out of the loop, and capability overhang all *increase* with increased access. The risk of fueling a race to the bottom *increases* when we assume a "winner takes all" dynamic. Only the risk of reinforcing bias *fluctuates* as access increases.

Increased access to foundation models also presents a range of opportunities, including those outlined earlier in this report. Proponents of increased access generally cite two main benefits:

- **Increased capacity for transparency, accountability, and reproducibility**. Increased access to foundation models and components allows researchers to scrutinize and validate each other's work, which could result in a culture of openness and rigorous scientific inquiry, red-teaming, and ultimately more secure systems.[117,118,119]
- **Technological democratization and promoting healthy competition.** Increased access to foundation models may democratize the technological ecosystem by lowering the technical barrier to entry for competitors, encouraging innovation, fostering diversity and inclusion, and decreasing the likelihood of monopolistic control over powerful models by only a handful of leading AI companies.[120]

While it is true that increased access to these models may encourage these outcomes—and in fact has done so in many cases—it is notable that the majority of identified opportunities are "downstream," and will thus require significant time and resources from users, AI labs, and regulatory agencies to be realized. At times, effectively leveraging AI technologies to realize these opportunities will require the restructuring of existing systems of incident management. On the contrary, the majority of the risks identified in the matrix are either entirely or in part "upstream" risks (functions of the models themselves and the labs that develop them), or the

[117] Stefan Larsson and Fredrik Heintz, "Transparency in Artificial Intelligence," *Internet Policy Review* 9, no. 2 (2020), https://policyreview.info/pdf/policyreview-2020-2-1469.pdf.

[118] Inioluwa Deborah Raji et al., "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020): 33-44, https://dl.acm.org/doi/abs/10.1145/3351095.3372873.

[119] Ana Lucic et al., "Reproducibility as a Mechanism for Teaching Fairness, Accountability, Confidentiality, and Transparency in Artificial Intelligence," *Proceedings of the AAAI Conference on Artificial Intelligence* 36 no. 11 (2020): 12792-800, https://ojs.aaai.org/index.php/AAAI/article/view/21558.

[120] "Deep Dive: AI Final Report," Open Source Initiative, accessed November 10, 2023, https://deepdive.opensource.org/wp-content/uploads/2023/02/Deep-Dive-AI-final-report.pdf.

result of “downstream” misuse that can be facilitated by a single actor or small group of actors, at times without the need for access to significant new resources.

Further, while increased access to foundation models may democratize the AI ecosystem and promote healthy competition at the technological and economic levels, thereby preventing monopolies over AI development, it is important to also consider the impact of risk, economic or otherwise, especially in extreme and catastrophic cases. Increasing access to foundation models is not the only way to prevent a monopoly in this space–this can also still be achieved by utilizing structured access, rate limiting, gating, and other technical mechanisms to reduce the possibility of harm. Regulatory actions can also impose liability on foundation model developers and users. The next phase of this project will work to delineate the most effective mechanisms to reduce risk. Finally, increasing unmediated access to these models may facilitate entire categories of risk that cannot be rolled back.

Today’s digital ecosystem is not yet broadly secure and sustainable, in many ways due to the lack of secure and safe design principles built into emerging technologies from the outset. Safety and the mitigation of harm should not be sacrificed solely in the name of rapid innovation, and lessons of the recent past in other areas of technological innovation, like social media, provide us with the opportunity to leverage AI for good.

There are a number of challenges to overcome in both managing and/or regulating AI, not least of which are jurisdictional issues. Allowing downloadable and fully open access to models will further complicate efforts to manage and/or regulate AI by diluting the potency of regulation targeting developer organizations and increasing the potential that users will have access to unregulated models. For example, effective regulation aiming to rein in access to AI foundation models would have to address open source models like Falcon 40B, developed by the United Arab Emirates’ Advanced Technology Research Council,[121] among others.

Finally, AI misuse will exert impact not only via digital systems, but also physical and social systems. In comparison to vulnerability discovery in a context like computer security, many forms of AI misuse are more difficult and costly to defend against.[122] While in the case of open source code, vulnerabilities may exist, and malicious actors may leverage these vulnerabilities for nefarious aims, AI exists not only as a component in products and services (like open source code), but also often as a product or service itself. Averting the risk that increased access to foundation models will facilitate the creation of deep fakes or widespread misinformation, for example, will require interventions not only in the technical capabilities

[121] Lisa Barrington, “Abu Dhabi makes its Falcon 40B AI model open source,” *Reuters*, May 25, 2023, https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/.

[122] Toby Shevlane and Allan Dafoe, “The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?” arXiv, January 9, 2020, https://arxiv.org/abs/2001.00463.

of models themselves, but also in the social systems in which they are deployed. The further upstream these risks can be diverted, the better.

Looking Ahead

Increasing access to AI foundation models is only one way to achieve many of the perceived benefits of AI. Many of the associated benefits may also be accomplished alongside technical mechanisms, regulatory structures, and voluntary agreements between leading AI labs.

We hope that this research catalyzes discussions about how best to reduce the risk of harm posed by increased access to AI foundation models. We look forward to gathering feedback about our analysis and continuing to engage constructively in this dialogue.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org