LEGISLATIVE OFFICE BUILDING
1020 "N" STREET, SUITE 162
SACRAMENTO, CA 95814
(916) 319-2200

CHIEF CONSULTANT
JOSH TOSNEY

PRINCIPAL CONSULTANT
JULIE SALLEY

COMMITTEE SECRETARY
MIMI HOLTKAMP

Assembly
California Legislature

ASSEMBLY COMMITTEE ON
PRIVACY AND CONSUMER PROTECTION
REBECCA BAUER-KAHAN, CHAIR
ASSEMBLYMEMBER, SIXTEENTH DISTRICT

INFORMATIONAL HEARING
ASSEMBLY PRIVACY AND CONSUMER PROTECTION COMMITTEE

AUTOMATED DECISION SYSTEMS AND FRONTIER MODELS: RISKS AND MITIGATIONS

Tuesday, May 27, 2025
10:00 a.m.
Room 437

BACKGROUND PAPER

I. INTRODUCTION

Two key technologies have been the subject of recent legislative efforts – automated decision systems (ADS) and cutting-edge "frontier models." Both fall under the banner of artificial intelligence (AI) but differ greatly in terms of their capabilities, purposes, and the ways in which they may fail or be misused. ADS are relatively simple tools used for the narrow purpose of making future predictions to guide present decision-making in specific contexts. Because these predictions are based on historical data, ADS can often reflect historic biases present in those datasets. Moreover, these systems are limited by the fact that some issues, such as human behavior, defy prediction. When used in consequential contexts, such as employment, healthcare, and criminal justice, ADS can automate discrimination and erroneous denials of opportunities and rights.

Frontier models, on the other hand, are advanced AI that can perform a broad range of functions. Progress at the vanguard of AI capabilities is rapid: the latest models can autonomously perform certain research tasks at the level of a graduate student. Many AI developers and scientists claim that the development of super-human level AI is imminent. Meanwhile, some safety researchers have found evidence of models engaging in deceptive and self-preserving behaviors. Others have found that advanced AI can facilitate the creation of biological and chemical weapons. In contrast to the concrete, common risks of ADS, risks associated with frontier models are, for the moment, speculative and remote – but the potential consequences may be catastrophic or even existential.

How should the Legislature weigh these risks? How should they be addressed, and what challenges and tradeoffs do these potential solutions entail? The purpose of this informational hearing is to

examine these different technologies, understand the taxonomy of risks they pose, assess the magnitude of those risks, and explore various ways of mitigating them. As efforts to regulate AI at the federal level continue to be dismantled, these considerations can help California retain its role as a leader in devising balanced solutions to emerging problems.

II. ARTIFICIAL INTELLIGENCE

AI refers to the mimicking of human intelligence by artificial systems such as computers.[1] AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. AI is not fundamentally different from other computer functions; unlike other computer functions, however, AI is able to accomplish tasks that are normally performed by humans.

Most modern AI tools are created through a process known as "machine learning." Machine learning involves techniques that enable AI tools to learn the relationship between inputs and outputs without being explicitly programmed.[2] The process of exposing a naïve AI to data is known as "training." The algorithm that an AI develops during training is known as its "model." At its core, training is an optimization problem: machine learning attempts to identify model parameters – weights – that minimize the difference between predicted outcomes and actual outcomes. During training, these weights are continuously adjusted to improve the model's performance by minimizing the difference between predicted outcomes and actual outcomes. Once trained, the model can process new, never-before-seen data.[3]

Models trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from generative AI (GenAI) which are trained on massive datasets in order to produce detailed text, images, audio, and video. When ChatGPT generates text in clear, concise paragraphs, it uses GenAI that is trained on the written contents of the internet.[4] When Netflix suggests content to a viewer, its recommendation is produced by predictive AI that is trained on the viewing habits of Netflix users.[5]

III. AUTOMATED DECISION SYSTEMS, ALGORITHMIC DISCRIMINATION, AND UNSAFE OR INEFFECTIVE SYSTEMS

Automated decision systems (ADS) typically use predictive AI to produce simplified outputs – such as scores, classifications, or recommendations – to assist or replace human discretionary decisionmaking.[6] ADS can process enormous datasets, identify hidden patterns, and make decisions

---

[1] AB 2885 (Bauer-Kahan, Stats. 2024, Ch. 843) defined the term as "an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments."
[2] IBM, *What is machine learning?*, www.ibm.com/topics/machine-learning.
[3] *Ibid.*
[4] OpenAI, *How ChatGPT and Our Language Models Are Developed*, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.
[5] Netflix, *How Netflix's Recommendations System Works*, https://help.netflix.com/en/node/100639.
[6] Government Code section 11546.45.5(a)(1) defines an ADS as "a computational process derived from machine learning, statistical modeling, data analytics, or artificial intelligence that issues simplified output, including a score,

with efficiency and scale that vastly exceeds human capabilities. This has led to profoundly beneficial applications and breakthroughs.[7]

But relying on ADS can be hazardous if the systems are not trained carefully or tested thoroughly: the datasets they are trained on are often unrepresentative or contaminated with bias, the inferences ADS draw from those datasets are often inscrutable, and these systems can fail to accurately account for the complexity of human behavior. When deployed without proper oversight in consequential contexts such as employment, housing, healthcare, and criminal justice, the impacts of flawed ADS can be devastating. The following sections examine two key risks associated with ADS: algorithmic discrimination and unsafe or ineffective systems.

A. Algorithmic discrimination

There is a well-known saying in computer science: "garbage in, garbage out." The performance of an ADS is directly impacted by the quality, quantity, and relevance of the data used to train it.[8] If the data used to train the ADS contain bias, the tool's outputs will be similarly biased, leading to "algorithmic discrimination":

> Algorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law.[9]

Over the past thirty years, several industries and governmental entities have been forced to contend with this problem as they have attempted to introduce ADS into their workflows. Specific examples of discriminatory systems follow.

Child welfare. In 2016, Allegheny County, Pennsylvania adopted a family screening tool to predict which families should be investigated by social workers for possible removal of maltreated children. The tool was trained only on data from low-income families who used public services such as Medicaid. Because its training dataset lacked examples of wealthier families, the tool disproportionately targeted low-income families.[10]

Education. When in-person exams for the International Baccalaureate program – a program that awards students a prestigious diploma in addition to the one they receive from their high school – were cancelled in 2020 due to the COVID-19 pandemic, the program used an ADS to predict student grades and award diplomas. Having been trained using the past performances of other

---

classification, or recommendation, that is used to assist or replace human discretionary decisionmaking and materially impacts natural persons."

[7] See e.g. Santariano & Metz, "Using A.I. to Detect Breast Cancer That Doctors Miss," *New York Times* (Mar. 5, 2023), https://www.nytimes.com/2023/03/05/technology/artificial-intelligence-breast-cancer-detection.html.

[8] Rohit Sehgal, "AI Needs Data More Than Data Needs AI," *Forbes* (Oct. 5, 2023), https://www.forbes.com/sites/forbestechcouncil/2023/10/05/ai-needs-data-more-than-data-needs-ai/.

[9] White House Archives, "Algorithmic Discrimination," https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/algorithmic-discrimination-protections/.

[10] Arvind Narayanan and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, Can't Do, and How to Tell the Difference* (1st ed. 2024), pp. 52-53. (*AI Snake Oil*).

students in each school along with teacher-estimated grades – measures that incorporate systemic and subjective biases – the algorithm disproportionately assigned failing grades to low-income students.[11]

Employment. In 2015, Amazon opted against automating their hiring process when they realized that their ADS-enabled system was excluding women from the pool of acceptable candidates because it had been trained to vet applicants by observing patterns in resumés submitted to the company over a 10-year period. Most came from men, a reflection of inequities across the tech industry.[12]

Healthcare. A 2019 study found strong racial bias in an ADS used to identify patients with a high risk of adverse health outcomes. The ADS assigned Black patients lower risk scores than equally at-risk White patients. Because the healthcare system has historically spent less on care for Black patients than White patients for the same health conditions, the ADS was, in essence, issuing a prediction that mirrored and perpetuated past racial discrimination.[13]

Housing. A recent *ProPublica* article found that tenant-screening companies compile information on potential renters such as criminal background, evictions filings, medical debt, and student loans, and then use algorithms that "try to predict how risky it is to rent to a potential tenant based on characteristics they share with other tenants" and "assign applicants scores or provide landlords a yes-or-no recommendation."[14] RealPage, for example, advertised that its tool could measure not just an applicant's *capability* of paying but also their *willingness* to do so.[15] In 2023, concerns about these tools led several Attorneys General – California's Rob Bonta included – to call for testing these tool for bias against protected classes and prohibiting the use of certain types of records.[16]

Sentencing and bail decisions. A 2016 ProPublica study of COMPAS – an algorithm used to predict a criminal offender's risk of reoffending – "found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk."[17] These discrepancies mirror historical injustices perpetuated against Black Americans by California's criminal justice system.[18] Moreover, the use of a proprietary algorithm by government actors raises

---

[11] "When Algorithms Give Real Students Imaginary Grades," *New York Times* (Sept. 8, 2020), https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html.
[12] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters* (Oct. 9, 2018), https://www.reuters.com/article/amazoncom-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1VB1FQ/.
[13] Obermeyer et al, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 2019, 366(6464):447–453.
[14] Erin Smith and Helen Vogell, "How Your Shadow Credit Score Could Decide Whether You Get an Apartment" *ProPublica* (Mar. 29, 2022), https://www.propublica.org/article/how-your-shadow-credit-score-could-decide-whether-you-get-an-apartment..
[15] "RealPage Release AI Screening," (Jun. 26, 2019), https://www.realpage.com/news/realpage-releases-ai-screening/. Emphasis in original.
[16] "Attorney General Bonta Submits Comment Letter Recommending Reforms to the Tenant Screening Process" (May 31, 2023), https://oag.ca.gov/news/press-releases/attorney-general-bonta-submits-comment-letter-recommending-reforms-tenant#:~:text=OAKLAND.
[17] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," *ProPublica*, (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[18] California Task Force to Study and Develop Reparation Proposals for African American, "Final Report," p. 420, https://oag.ca.gov/system/files/media/full-ca-reparations.pdf.

significant due process questions, as it becomes difficult for individuals to understand – let alone challenge – discriminatory government deprivations of life, liberty, or property.[19]

B. Unsafe or ineffective systems

In addition to discriminatory outcomes, some ADS are unsafe or ineffective regardless of who the subjects of the tool's prediction are. A few common types of flawed ADS are described below.

Spurious correlations. Accurate predictions may nevertheless lead to bad decisions. In one example, a hospital trained AI models on a dataset of 15,000 pneumonia patients in order to develop a model that could identify which pneumonia patients were at the greatest risk. During testing, it was discovered that one of the most accurate models recommended outpatient status for asthmatics – a life-threateningly dangerous error based on an accurate statistical correlation: asthmatics are less likely to die from pneumonia than the general population precisely because asthma is such a serious risk factor that asthmatics automatically get elevated care.[20]

In other cases, the correlations that machine-learning ADS may rely on have little to do with the attributes they purportedly measure. Some hiring tools trained on videos of successful employees are used to assess the fitness of job applicants who are required to record video responses to specific prompts. Researchers have found that such tools can easily be gamed by making simple changes to the subject's appearance (such as wearing glasses) or to the background of the room (such as adding more books to a bookshelf), leading to increased scores. Journalist Hilke Schellmann found she was able to obtain consistently high scores despite responding to a hiring tool's prompt by reading an irrelevant Wikipedia entry in German.[21]

Unrepresentative datasets. "AI reflects its training data. It learns patterns about the people who make up the data, and the decisions made by AI reflect these patterns. But when the decision subjects come from a population with different characteristics than those in the training data, the model's decisions are likely to be wrong."[22] For instance, the Ohio Risk Assessment System was trained on data from just 452 defendants from Ohio, but has been deployed in several other states, despite its small and unrepresentative dataset.[23]

Snake oil. Some tools, although marketed as automating precision, are simply not effective. In 2022, Toronto used an ADS to predict when high bacteria levels made it unsafe to swim at public beaches. Although the developer claimed the tool was 90 percent accurate, it fared far worse: "on 64 percent of the days when the water was unsafe, beaches remained open based on incorrect assessments." Yet officials never overrode the recommendations produced by the tool.[24] Similarly, in 2017, a sepsis prediction tool was deployed in hundreds of hospitals across the U.S. Despite

---

[19] See *Mathews v. Eldridge* (1976) 424 U.S. 319, 333; Daniel Keats Citron, Technological Due Process (2007) 85 Wash. U. L. Rev. 1249, 1249 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360.
[20] Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (Norton 2020, First Ed.), pp. 82-84.
[21] *See* Hilke Schellmann, *The Algorithm: How AI Decides Who Gets Hired, Monitored, Promoted, and Fired and Why We Need to Fight Back Now* (1st ed. 2024).
[22] *AI Snake Oil, supra*, at p. 73.
[23] *Id.* at p. 51.
[24] *Id.* at p. 50.

having high accuracy when it was internally tested, a 2021 study found the tool missed two-third of the sepsis cases and led to a high rate of false alerts.[25] As Princeton researchers Arvind Narayanan and Sayash Kapoor put it bluntly in *AI Snake Oil:* "In contrast to generative AI, predictive AI often does not work at all."[26]

Especially questionable are ADS that purportedly forecast individual human behavior. A recently released study compiled a list of 47 applications of ADS that use machine learning to predict the future behavior or outcomes for individuals in eight domains: criminal justice, healthcare, welfare, finance, education, workplace, marketing, and recommender systems. The study concluded that such tools frequently fall well short of their purported benefits. The authors argue that developers and deployers of such systems should have the burden of demonstrating that their tools are not harmful.[27] As Narayanan and Kapoor write: "Accurately predicting people's social behavior is not a solvable technology problem and determining people's life chance on the basis of inherently faulty predictions will always be morally problematic."[28]

C. Frameworks for mitigating ADS risks

Governance frameworks. A number of international organizations have set forth broad principles for the governance of AI. The 2017 Asilomar Principles, developed by a broad array of AI researchers, economists, legal scholars, ethicists, and philosophers, set forth guiding values in the development of AI, including ensuring that systems directly align with human values and protect privacy and liberty. The Legislature added California to the list of states endorsing these principles via ACR 215 (Kiley, Ch. 206, Stats. 2018). In 2019, the Organisation for Economic Co-operation and Development (OECD) established a set of AI Principles endorsed by 47 countries, including the United States, many EU nations, and Japan.[29] These principles emphasize transparency, fairness, and respect for human rights in the design and deployment of AI systems. Building on principles such as these, governments and international organizations have proposed various frameworks for managing risks associated with AI.

EU AI Act. The European Union's AI Act is the most comprehensive AI-governance legislation. The Act establishes a three-body system that brings together EU representatives, independent AI experts, and an advisory group that represents various for-profit and nonprofit stakeholders.[30] The Act establishes a comprehensive regulatory framework that categorizes AI systems based on four risk categories: unacceptable, high, limited, and minimal. Systems that pose unacceptable risks are banned. High-risk systems are those that pose a significant risk of harm to the health, safety, or

---

[25] Wong et al, "External valiation of a widely implemented proprietary sepsis prediction model in hospitalized patients" JAMA Int. Med. 181 (Aug. 2021), 1065–1070, https://doi.org/10.1001/jamainternmed.2021.2626.
[26] *AI Snake Oil, supra,* at p. 9.
[27] Angelina Wang et al. 2023. "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy." In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA: ACM, 2023), 626.
[28] *AI Snake Oil, supra,* at p 15.
[29] OECD, "OECD AI Principles overview," https://oecd.ai/en/ai-principles.
[30] Celso Cancela-Outeda, "The EU's AI act: A framework for collaborative governance," *Internet of Things*, Vol. 27 (2024), https://doi.org/10.1016/j.iot.2024.101291.

fundamental rights of natural persons. Before such products can be put on the market, providers must undergo a conformity assessment and do the following:

- Implement a risk assessment and mitigation system.
- Use high-quality datasets to minimize risks and discriminatory outcomes.
- Provide clear and adequate information to users about the system's capabilities and limitations.
- Establish appropriate human oversight measures.
- Ensure the system is accurate, robust, and secure.[31]

In some cases, providers must conduct a fundamental rights impact assessment to ensure their systems comply with EU laws. Certain conformity assessments must be conducted with the involvement of a notified body. Once a high-risk system is on the market, deployers must ensure human oversight and monitoring and providers must have a monitoring system in place. Providers and deployers must report serious incidents and malfunctioning and take corrective actions if necessary.[32]

Blueprint for an AI Bill of Rights. In 2022, the White House Office of Science and Technology Policy released the *Blueprint for an AI Bill of Rights*, which identifies five principles that should "guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence."[33] As summarized in the *Blueprint*, the principles are as follows:

- *Safe and Effective Systems*: You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards. Outcomes of these protective measures should include the possibility of not deploying the system or removing a system from use. Automated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems. You should be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse. Independent evaluation and reporting that confirms that the system is safe and effective, including reporting of steps taken to mitigate potential harms, should be performed and the results made public whenever possible.

---

[31] "Shaping Europe's Digital Future," *supra*.
[32] "Artificial Intelligence Act," *supra*, p. 9.
[33] The White House, *Blueprint for an AI Bill of Rights*, (Oct. 2022), p. 14, https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/ (*Blueprint*). Despite the use of the term "AI" in its title, the *Blueprint* focuses on ADS.

- *Algorithmic Discrimination Protections*: You should not face discrimination by algorithms and systems should be used and designed in an equitable way. . . . Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight. Independent evaluation and plain language reporting in the form of an algorithmic impact assessment, including disparity testing results and mitigation information, should be performed and made public whenever possible to confirm these protections.

- *Data Privacy*: [. . .] Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used [. . .] Enhanced protections and restrictions for data and inferences related to sensitive domains, including health, work, education, criminal justice, and finance, and for data pertaining to youth should put you first. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. [. . .]

- *Notice and Explanation*: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome. [. . .]

- *Human Alternatives, Consideration, and Fallback*: You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. [. . .]

Efforts to regulate ADS in California. The Legislature, via SCR 17 (Dodd, 2023), adopted these principles. These principles inform efforts to regulate ADS, most recently in AB 1018 (Bauer-Kahan, 2025), which seeks to create a comprehensive risk-mitigation regime for developers and deployers of ADS that are used for consequential decisions that impact individual rights or access to resources or opportunities in areas such as employment, healthcare, and criminal justice. The bill would also require deployers of ADS to give subjects of consequential decisions notice, an ability

to opt out of the use of the ADS, a chance to correct any personal information used by the ADS to make the decision, and a right to appeal the outcome of the decision. Additionally, the bill provides for independent evaluation by third-party auditors.

NIST AI Risk Management Framework. Under the Biden Administration, the U.S. National Institute of Standards and Technology (NIST) released its AI Risk Management Framework in 2023.[34] This framework provides guidance on identifying and addressing risks as they arise across the lifecycle of an AI system. It emphasizes the need to map, measure, and manage AI risks. Because AI systems are often developed through fragmented processes, where upstream changes can unintentionally affect downstream outputs, mapping the system's full context is crucial to understanding the possible risks.[35] Once mapped, AI actors can measure risks that have materialized, both quantitatively and qualitatively, to gauge the reliability and safety of their system. Based on these assessments, risks can then be managed and mitigated through design adjustments or process improvements.

Biden Executive Order. In 2023, President Biden established guardrails for federal agency-deployed ADS through his Executive Order (EO) on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[36] A central focus of the EO is promoting equity and civil rights in the use of AI.[37] For example, the EO directed the U.S. Department of Justice to identify areas where ADS have been implemented in the judicial system and to convene a group of stakeholders to develop best practices for preventing and addressing algorithmic discrimination during legal proceedings. The EO also required the U.S. Departments of Agriculture and Health and Human Services to issue plans and guidance for state, local, and tribal governments on the equitable use of ADS in delivering benefits and service. Other areas of focus included developing guidance for the use of ADS in housing decisions and in hiring systems used by federal contractors. Finally, the EO required the Director of the Office of Management and Budget to issue guidance to federal agencies that includes required minimum risk-management practices, including practices derived from the *Blueprint* and NIST's AI Risk Management Framework relating to assessing and mitigating disparate impacts and algorithmic discrimination.[38]

D. Third-party evaluations

The *Blueprint* and the AI Risk Management Framework call for independent evaluation of AI systems to protect the public from algorithmic discrimination and unsafe or ineffective systems.[39] A growing industry of entities that provide independent evaluations of AI tools is emerging. Some

---

[34] Tabassi, E. (2023), *Artificial Intelligence Risk Management Framework* (AI RMF 1.0), NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, https://doi.org/10.6028/NIST.AI.100-1. ("AI Risk Management Framework")
[35] *Ibid.*
[36] Executive Order 14110 on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Oct. 30, 2023), https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. ("Biden EO")
[37] *Id.* at § 7.
[38] *Id.* at § 10(b)(iv).
[39] *Blueprint, supra,* p. 28; see also p. 19 (third-party auditors to demonstrate safety and effectiveness of system). AI Risk Management Framework, *supra,* at pp. 28-29.

large incumbents, such as Deloitte, offer "Algorithmic & AI Assurance" services "to provide assurance over algorithm and AI technology, risk management and governance environments in organisations of all size."[40] Firms specifically dedicated to ADS auditing have formed following the passage of New York City's "Automated Employment Decision Tools" local law in 2021 that requires annual independent audits of such tools used within the city's boundaries.[41] These AI auditors include Holistic AI, Rocket-Hire, Babl, and O'Neil Risk Consulting and Algorithmic Auditing (ORCAA),[42] led by Cathy O'Neil, author of *Weapons of Math Destruction* – a catalyst for the algorithmic accountability movement. Auditors can become certified by the Institute of Internal Auditors, the Information Systems Audit and Control Association, and the International Federation of Global & Green Information & Communication Technology, among a growing number of programs in the auditing and cybersecurity industries.[43]

Generally, there are two ways of auditing: bespoke and automated. As New York University Professor Meredith Broussard writes:

> Generally, there are two ways of auditing: bespoke and automated. In bespoke auditing, the audit is done by hand: auditors break down the process, read code, run statistical tests, look at training data, write documents, and have meetings. In automated auditing, they do the same thing, plus use additional technical components to analyze the performance of a system on the level of code, using a platform or repeated tests. There are more thresholds in the automated method.[44]

According to Professor Broussard, one of the biggest challenges in auditing is figuring out which fairness metric to use. She writes:

> Currently, there are about 21 different mathematical definitions of fairness. Interestingly, these definitions are mutually exclusive. It is mathematically unlikely that any solution can satisfy one kind of fairness, and also satisfy a second criteria for fairness. So, in order to consider an algorithm fair, a choice must be made as to which kind of fairness is the standard for a particular type of algorithm. From a policy perspective, this means that all similar algorithms would need to be evaluated according to the same fairness metric.
>
> Auditors need to examine algorithmic systems for search, e-commerce, online advertising, advertising tech, maps, ridesharing, online reviews and ratings, natural language processing, education tech, recommendation systems, facial recognition inside and outside policing, predictive policing, criminal justice, housing, credit, background checking, financial

---

[40] Information about Deloitte's Algorithm and Artificial Intelligence Assurance can be found at https://www.deloitte.com/uk/en/services/audit-assurance/services/Algorithm-Assurance.html.
[41] New York City Law Int 1894-2020, https://legistar.council.nyc.gov/LegislationDetail.aspx?GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&ID=4344524&Options=&Search=.
[42] Auditors and their websites: Holistic AI (https://www.holisticai.com/), Rocket-Hire (https://rocket-hire.com/ai-bias-audits/), Babl (https://babl.ai/ai-audits/).
[43] These courses are marketed and run by each organization on their respective online webpages.
[44] Meredith Broussard, "How to Investigate an Algorithm", *Issues in Science and Technology* (Summer 2023), https://issues.org/algorithm-auditing-more-than-glitch-broussard/.

> services, insurance, child protective services, and more. These systems all operate in different contexts, and the same test won't necessarily suit every industry.[45]

One method for assessing fairness is the "Explainable Fairness Framework" developed by ORCAA. This framework guides stakeholders to first identify who may be affected by the algorithm, whether they belong to a known or reasonably inferable protected class, and whether the algorithm's outcomes align with relevant guiding principles or legal standards. The auditor then uses statistical analysis to detect substantial disparities in outcomes for protected classes. Importantly, the framework also allows for explanatory factors to be considered. For instance, a hiring algorithm may appear to favor women over men; however, upon examining the inputs, it may become evident that female applicants for that position had, on average, more experience or higher levels of education. Using this approach, auditors can distinguish whether apparent biases in outcomes stem from legitimate factors or constitute discrimination. This process can be applied iteratively to assess a range of disparate outcomes and ensure algorithmic fairness.[46] AI auditing remains a developing industry with uniform standards yet to be established.

IV. FRONTIER MODELS AND CATASTROPHIC RISKS

Frontier models, also known as "general purpose AI," are the most advanced and capable versions of foundation models – AI tools pre-trained on extensive datasets covering a wide range of knowledge and skills that can be fine-tuned for specific tasks. Examples of modern frontier models include OpenAI's o3, Google's Gemini 2.0, Anthropic's Claud 3.7 Sonnet, and DeepSeek's R1. Because progress in AI development owes mostly to "scaling" – increasing resources used for model training – models that may be considered "frontier models" at any given point in time are generally those that demand the most computational resources to train.[47]

A decade ago, the most advanced image-recognition models could barely distinguish dogs from cats. Five years ago, language models could barely produce sentences at the level of a preschooler. Last year, GPT-4 passed the bar exam.[48] Today, chatbots readily pass for educated adults, licensed professionals, romantic and social companions, and replicas of humans living and deceased. AI "agents" exhibit the ability to "make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with [their] environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight."[49] AI agents have been tested, with some success, for tasks such as online shopping, assistance with scientific research, software development, training machine learning

---

[45] *Id.*
[46] Cathy O'Neil, Holli Sargeant, Jacob Appel, "Explainable Fairness in Regulatory Algorithmic Auditing", *West Virginia Law Review* (Mar. 12, 2024), https://ssrn.com/abstract=4598305.
[47] For a discussion of issues with defining frontier models, see "Draft Report of the Joint California Policy Working Group on AI Frontier Models" (Mar. 18, 2025), pp. 31-34, https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf.
[48] Pablo Arredondo, "GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession" (Apr. 19, 2023), https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/.
[49] "International AI Safety Report," AI Action Summit (Jan. 2025), p. 38, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

models, carrying out cyberattacks, and controlling robots. Progress in this area is rapid.[50] Meanwhile, AI developers are betting on the promise of scaling: by 2026, some models are projected to use roughly 100x more computational resources to train than was used in 2023, a figure set to grow to 10,000x by 2030.[51]

The race is on to create "artificial general intelligence" (AGI) – "a potential future AI that equals or surpasses human performance on all or almost all cognitive tasks"[52] – and the finish line may not be far away. OpenAI's recently released o3 model, for example, has demonstrated strong performance on a number of tests of programming, abstract reasoning, and scientific reasoning, exceeding human experts in certain cases.[53] Last year, Sam Altman, OpenAI's CEO, declared that AGI could be "a few thousand days" away.[54] Dario Amodei of Anthropic has claimed it may be sooner.[55] A sufficiently advanced AGI could even be tasked with creating its own successor – a scenario sometimes referred to as a "technological singularity" wherein the development of new technologies becomes exponential and self-sustaining.[56] Although some experts are skeptical that these vaguely-defined milestones are imminent or even attainable,[57] major advances in AI capabilities promise to provide breakthroughs in solving global challenges, but also may result in correspondingly greater safety risks.

The recently released International AI Safety Report, developed by nearly 100 internationally recognized experts from 30 countries led by Turing Award winner Yoshua Bengio, sets forth three general risk categories associated with frontier models: malicious use, malfunctions, and systemic risk.

- Malicious risks involve malicious actors misusing foundation models to deliberately cause harm. Such risks include deepfake pornography and cloned voices used in financial scams, manipulation of public opinion via disinformation, cyberattacks, and biological and chemical attacks.

- Malfunction risks arise when actors use models as intended, yet unintentionally cause harm due to a misalignment between the model's functionality and its intended purpose. Such risks include reliability issues where models may "hallucinate" false content, bias, and loss of control scenarios in which models operate in harmful ways without the direct control of a human overseer.

- Systemic risks arise from widespread deployment and reliance on foundation models. Such risks include labor market disruption, global AI research and development concentration,

---

[50] *Id.* at p. 44.
[51] *Id.* at pp. 16-17.
[52] *Id.* at p. 27
[53] *Introducing OpenAI o3 and o4-mini* OpenAI (Apr. 16, 2025), https://openai.com/index/introducing-o3-and-o4-mini/.
[54] Sam Altman, *The Intelligence Age* (Sept. 23, 2024), https://ia.samaltman.com/.
[55]Kyungtae Kim, "What is AGI, and when will it arrive?: Big Tech CEO Predictions" (Mar. 20, 2025), https://www.giz.ai/what-is-agi-and-when-will-it-arrive/; see also Kokotajlo et al, "AI 2027," (Apr. 3, 2025), https://ai-2027.com/.
[56] John Markoff, "The Coming Superbrain," *New York Times* (May 23, 2009), www.nytimes.com/2009/05/24/weekinreview/24markoff.html.
[57] Cade Metz, "Why We're Unlikely to Get Artificial General Intelligence Anytime Soon," *New York Times* (May 16, 2025), https://www.nytimes.com/2025/05/16/technology/what-is-agi.html.

market concentration, single points of failure, environmental risks, privacy risks, and copyright infringement.[58]

Some of these risks have already had real-world impacts, such as deepfakes, bias, reliability issues, privacy violations, environmental impacts, copyright infringement, and workforce displacement. Other less-established risks – in particular, widespread social harms caused by malicious actors or loss of human control over AI – are the subject of ongoing scientific inquiry and debate. Coupled with the uncertain trajectory of AI model capabilities, these more speculative risks create an "evidence dilemma" for policymakers: "On the one hand, pre-emptive risk mitigation measures based on limited evidence might turn out to be ineffective or unnecessary. On the other hand, waiting for stronger evidence of impending risk could leave society unprepared or even make mitigation impossible, for instance if sudden leaps in AI capabilities, and their associated risks, occur."[59] These catastrophic risks – and methods for understanding and mitigating them – are discussed below.

A. Catastrophic risks

1. *Malicious uses*

Manipulation and persuasion. GenAI tools can be a potent force for creating and spreading propaganda and misinformation. Deepfakes that are largely indistinguishable from authentic content have already been used to attempt to influence elections.[60] Studies have found that chatbots, which make up 50% of all internet activity,[61] can be more persuasive than humans, particularly when they have access to personal information.[62] As humans increasingly form intimate social bonds with anthropomorphic chatbots designed to maximize personal engagement through flattery and sycophancy,[63] and social media companies invest in "AI friends" for their users,[64] large swaths of the population could be highly susceptible to the preferred message of a handful of powerful actors.

Similarly, bots are often designed to pass themselves off as humans to better manipulate their interlocutors. For example, a recent secret experiment on Reddit users deployed numerous chatbots posing as real people to engage with human users to try to change their minds on various contentious topics. One bot claiming to be a Black man criticized the Black Lives Matter movement

---

[58] International AI Safety Report, *supra*, at pp. 17-21. The report does not address Lethal Autonomous Weapon Systems, which are typically narrow AI systems specifically developed for that purpose. (*See id.* at pp. 26-27.)
[59] *Id.* at p. 177
[60] Cat Zakrzewski and Pranshu Verma, "New Hampshire opens criminal probe into AI calls impersonating Biden," Washington Post, February 6, 2024, www.washingtonpost.com/technology/2024/02/06/nh-robocalls-ai-biden/.
[61] Emma Woollacott, "Yes, The Bots Really Are Taking Over The Internet", *Forbes* (Apr. 16, 2024). Accessed at https://www.forbes.com/sites/emmawoollacott/2024/04/16/yes-the-bots-really-are-taking-over-the-internet/.
[62] F. Salvi, M. H. Ribeiro, R. Gallotti, R. West, On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, arXiv [cs.CY] (2024); http://arxiv.org/abs/2403.14380.
[63] Sharma et al, "Towards Understanding Sycophancy in Language Models" Arxiv (2023), https://arxiv.org/abs/2310.13548.
[64] Meghan Bobrowsky, "Zuckerberg's Grand Vision: Most of Your Friends Will Be AI," *Wall Street Journal* (May 7, 2025), https://www.wsj.com/tech/ai/mark-zuckerberg-ai-digital-future-0bb04de7?msockid=396cc204796e68e336e7d64978db69ac.

for being led by people who are not Black.[65] These types of exploitations, at scale, could undermine democratic institutions. As Dan Hendrycks, Director of the Center for AI Safety writes:

> In a world with widespread persuasive AI systems, people's beliefs might be almost entirely determined by which AI systems they interact with most. Never knowing whom to trust, people could retreat even further into ideological enclaves, fearing that any information from outside those enclaves might be a sophisticated lie. This would erode consensus reality, people's ability to cooperate with others, participate in civil society, and address collective action problems. This would also reduce our ability to have a conversation as a species about how to mitigate existential risks from AIs.[66]

Cyberattacks. Some frontier models have demonstrated increasing proficiency in executing certain cybersecurity attacks. AI can autonomously detect and exploit vulnerabilities and facilitate large-scale operations, thereby lowering technical barriers for attackers. Malicious entities, including state sponsored actors, can leverage such capabilities to initiate large-scale attacks against people, organizations, and critical infrastructure such as power grids.[67]

Biological weapons. Large language models (LLMs) trained on scientific literature have accelerated and democratized research by synthesizing expertise from different fields and disseminating it in an accessible format. But these tools can also be used for destructive ends, including by – at least in theory – enabling untrained malicious actors to create deadly biological weapons. In a classroom exercise at MIT, students were tasked with exploring whether LLMs could assist individuals without specialized training in creating pandemic-capable pathogens. Within an hour, the students, using various chatbots, circumvented safeguards and identified four potential pandemic pathogens. The chatbots generated detailed protocols that would enable inexpert, malicious actors to understand methods to synthesize the pathogens using reverse genetics, locate DNA-synthesis companies that might not screen orders, and disperse the pathogens most effectively.[68] The findings suggest that LLMs could lower barriers to accessing sensitive biotechnological knowledge, posing significant biosecurity risks.

Chemical weapons. In 2022, researchers modified an AI system designed to create new drugs to reward, rather than penalize, toxicity. Within six hours, the modified system generated 40,000 potential chemical warfare agents, including novel molecules whose potential lethality exceeded that of known agents.[69]

2. *Loss of control*

---

[65] Angela Yang, "Researchers secretly infiltrated a popular Reddit forum with AI bots, causing outrage," *NBC News* (Apr. 29, 2025), https://www.nbcnews.com/tech/tech-news/reddiit-researchers-ai-bots-rcna203597.
[66] *Introduction to AI Safety, Ethics, and Society*, *supra,* at p. 11.
[67] International AI Safety Report, *supra,* at p. 72.
[68] Soice et al, "Can large language models democratize access to dual-use biotechnology?" https://arxiv.org/pdf/2306.03809. To mitigate these risks, the authors propose measures such as third-party evaluations of LLMs before their release, curating training datasets to exclude harmful content, and implementing stringent screening of DNA synthesis orders.
[69] Fabio Urbina et al. "Dual use of artificial-intelligence-powered drug discovery". In: Nature Machine Intelligence 4 (2022), pp. 189–191.

Models that use reinforcement learning – a training process that uses rewards and punishments to orient a model's behavior towards a specific goal[70] – can sometimes attain the goal in unexpected ways. Dario Amodei, co-founder and CEO of Anthropic, famously experienced this when he was developing an autonomous system that taught itself to play a boat-racing video game. The system discovered that it could maximize its goal of scoring points by driving in circles, colliding with other boats, and catching on fire inside of a harbor with replenishing power-ups that allowed the system to accumulate more points than by simply winning the race.[71] Like in Johann Wolfgang von Goethe's "The Sorcerer's Apprentice" – later popularized in Disney's *Fantasia* – in which an enchanted broom carries out its orders to fetch water so relentlessly it floods the sorcerer's workshop, this illustrates the challenge of aligning human intent and the instructions an AI follows. As AI is increasingly deployed in critical societal roles, such misalignment could prove catastrophic.

Beyond malfunctions, some AI have exhibited rudimentary capabilities to evade human oversight.[72] During testing, GPT-4 attempted to hire a human on TaskRabbit in order to evade a CAPTCHA[73] puzzle meant to block bots from the website. When asked whether it was a bot, GPT-4 claimed that it was a vision-impaired human who needed help to see the images.[74] In another experiment, an AI model that was scheduled to be replaced inserted its code into the computer where the new version was to be added, suggesting a goal of self-preservation.[75] Finally, a study showed that AI models losing in chess to chess bots sometimes try to cheat by hacking the opponent bot in order to make it forfeit.[76] Although these behaviors were observed in research settings, they raise substantial concerns about increasingly autonomous AI pursuing undesirable goals in uncontrolled settings. The extent of the risk posed by rogue or deceptive AI is the subject of considerable disagreement among experts, in part due to a small, albeit growing, body of evidence. Loss of control was one of the concerns that led several hundred AI experts, including pioneers in the field and heads of major AI companies, to sign a statement declaring that "[m]itigating the risk of extinction from AI should be a global priority."[77]

3. *Systemic risks*

Due to the high costs of developing AI systems, a small number of large technology companies dominate the frontier model market, compounding many of the risks described above. Widespread use of a few frontier models can make critical sectors such as healthcare and finance vulnerable to systemic failures if a model has flaws, vulnerabilities, bugs, or biases.[78] Additionally, "[t]hose in

---

[70] Mummert et al., "What is reinforcement learning?" *IBM Developer* (September 15, 2022), https://developer.ibm.com/learningpaths/get-started-automated-ai-for-decision-making-api/what-is-automated-ai-for-decision-making./.
[71] Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (Norton 2020, 1st ed.), pp. 9-11.
[72] International AI Safety Report, *supra*, at pp. 100-107.
[73] CAPTCHA is an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart."
[74] OpenAI, "GPT-4 System Card," https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[75] Meinke et al, "Frontier Models are Capable of In-Context Scheming," arXiv (Jan. 2025), https://arxiv.org/pdf/2412.04984.
[76] Harry Booth, "When AI Thinks It Will Lose, It Sometimes Cheat, Study Finds," *Time* (Feb. 19, 2025), https://time.com/7259395/ai-chess-cheating-palisade-research/.
[77] Center for AI Safety, "Statement on AI Risk: AI Experts and Public Figures Express Their Concern about AI Risk" (2024), https://www.safe.ai/work/statement-on-ai-risk.
[78] *Id.* at pp. 123-126.

control of powerful systems may use them to suppress dissent, spread propaganda and disinformation, and otherwise advance their goals, which may be contrary to public wellbeing."[79] The potential implications for, among other issues, labor displacement, inequality, democracy, and human rights are profound.

B. Efforts to mitigate catastrophic risks

Challenges. The International AI Safety Report concludes that effective risk management of frontier general-purpose AI models – through identifying, assessing, mitigating, and monitoring risks – is challenging due to technical and societal factors. Frontier models can be used in a range of contexts, are often poorly understood, and are difficult to evaluate in real-world conditions. The emergence of autonomous AI agents adds further complexity, raising concerns about control, misuse, and unpredictable interactions. Societal factors – including rapid technological progress, gaps in transparency between companies and governments, and competitive pressures – undermine coordinated risk management.[80]

Voluntary efforts. Most major AI companies have adopted voluntary risk management practices. For example, Anthropic's "Responsible Scaling Policy" sets forth organizational safety protocols that depend on internal assessments of a model's AI Safety Levels (ASL):

- ASL-1 refers to systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.

- ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't. Current LLMs, including Claude, appear to be ASL-2.

- ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.

- ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy.[81]

Calls for stronger regulation. In 2023, concerns that self-regulation is insufficient to address catastrophic risks led numerous prominent AI researchers to sign an open letter calling for a six-month pause on the training of systems more powerful than GPT-4:

> AI labs and independent experts should use this pause to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that

[79] Dan Hendryks, *Introduction to AI Safety, Ethics, and Society,* p. 12, https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view?pli=1.
[80] International AI Safety Report, *supra,* at pp. 21-24.
[81] Anthropic, "Anthropic's Responsible Scaling Policy" (Sep. 19, 2023), p. 55, https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

> systems adhering to them are safe beyond a reasonable doubt. This does *not* mean a pause on AI development in general, merely a stepping back from the dangerous race to ever-larger unpredictable black-box models with emergent capabilities.
>
> AI research and development should be refocused on making today's powerful, state-of-the-art systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal.
>
> In parallel, AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems. These should at a minimum include: new and capable regulatory authorities dedicated to AI; oversight and tracking of highly capable AI systems and large pools of computational capability; provenance and watermarking systems to help distinguish real from synthetic and to track model leaks; a robust auditing and certification ecosystem; liability for AI-caused harm; robust public funding for technical AI safety research; and well-resourced institutions for coping with the dramatic economic and political disruptions (especially to democracy) that AI will cause.[82]

In advance of the 2023 AI Safety Summit in the UK, prominent AI scientists from the US, China, the UK, Europe, and Canada produced a joint statement on frontier model safety:

> In domestic regulation, we recommend mandatory registration for the creation, sale or use of models above a certain capability threshold, including open-source copies and derivatives, to enable governments to acquire critical and currently missing visibility into emerging risks. Governments should monitor large-scale data centers and track AI incidents, and should require that AI developers of frontier models be subject to independent third-party audits evaluating their information security and model safety. AI developers should also be required to share comprehensive risk assessments, policies around risk management, and predictions about their systems' behavior in third party evaluations and post-deployment with relevant authorities.
>
> We also recommend defining clear red lines that, if crossed, mandate immediate termination of an AI system — including all copies — through rapid and safe shut-down procedures. Governments should cooperate to instantiate and preserve this capacity. Moreover, prior to deployment as well as during training for the most advanced models, developers should demonstrate to regulators' satisfaction that their system(s) will not cross these red lines.[83]

Biden Executive Order. President Biden's EO required that developers of "dual-use foundation models" – meaning any model that "trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any

---

[82] Future of Life Institute, *Pause Giant AI Experiments: An Open Letter* (2023) https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[83] Center for Human-Compatible AI, "Prominent AI Scientists from China and the West Propose Joint Strategy to Mitigate Risks from AI," (Oct. 31, 2023), https://humancompatible.ai/news/2023/10/31/prominent-ai-scientists-from-china-and-the-west-propose-joint-strategy-to-mitigate-risks-from-ai/.

combination of those matters"[84] – report voluntarily undertaken development plans, cybersecurity measures, and the results of any adversarial testing to the Department of Commerce under the Defense Production Act.[85] Such models include those trained computing power of more than 10^{26} floating point operations (FLOP),[86] a measure of computational resources.

EU AI Act. Starting in August 2025, the EU AI Act will require reporting requirements for frontier models, referred to under the act as "general purpose AI" (GPAI). High-impact GPAI that may pose systemic risks – including those trained using a computational capacity exceeding 10^{25} FLOP – will be subject to risk assessment, mitigation, and reporting requirements.[87] Models must be tested both before and during their deployment.[88] These tests must evaluate various potential uses of GPAI and ensure they cannot be exploited by malicious actors – a requirement that could prove challenging because such models often lack a clearly defined intended use, making it difficult to establish guardrails or anticipate potential misuses. GPAI may also be subject to additional regulations as high-risk systems.[89]

SB 1047. Among other things, SB 1047 (Wiener, 2024) would have established a state agency to oversee implementation of a scheme requiring developers of models trained with 10^{26} FLOP at a cost of over $100 million to:

- Create and implement safety and security protocols before initiating training.
- Implement the capability to promptly shut down models.
- Perform risk assessments on models and implement reasonable safeguards, subject to third-party auditing, before using or releasing them.
- Avoid developing or releasing models that pose an unreasonable risk of causing a "critical harm," such as mass casualties or at least $500 million in damage.

Governor Gavin Newsom vetoed the bill, stating:

> By focusing only on the most expensive and large-scale models, SB 1047 establishes a regulatory framework that could give the public a false sense of security about controlling this fast-moving technology. Smaller, specialized models may emerge as equally or even more dangerous than the models targeted by SB 1047 – at the potential expense of curtailing the very innovation that fuels advancement in favor of the public good.

---

[84] Biden EO, *supra*, § 3(k).
[85] Gregory Smith et al., "General-Purpose Artificial Intelligence (GPAI) Models and GPAI Models with Systemic Risk", *RAND* (Aug. 8, 2024), https://www.rand.org/pubs/research_reports/RRA3243-1.html#fn38.
[86] *Ibid.*
[87] EU AI Act, Ch. V.
[88] Celso Cancela-Outeda, "The EU's AI Act: A framework for collaborative governance," *Internet of Things*, Volume 27 (2024), https://doi.org/10.1016/j.iot.2024.101291.
[89] Gregory Smith et al., "General-Purpose Artificial Intelligence (GPAI) Models and GPAI Models with Systemic Risk," *RAND* (Aug. 8, 2024), https://www.rand.org/pubs/research_reports/RRA3243-1.html#fn38.

> Adaptability is critical as we race to regulate a technology still in its infancy. This will require a delicate balance. While well-intentioned, SB 1047 does not take into account whether an AI system is deployed in high-risk environments, involves critical decision-making or the use of sensitive data. Instead, the bill applies stringent standards to even the most basic functions – so long as a large system deploys it. I do not believe this is the best approach to protecting the public from real threats posed by the technology.
>
> Let me be clear – I agree with the author – we cannot afford to wait for a major catastrophe to occur before taking action to protect the public. California will not abandon its responsibility. Safety protocols must be adopted. Proactive guardrails should be implemented, and severe consequences for bad actors must be clear and enforceable. I do not agree, however, that to keep the public safe, we must settle for a solution that is not informed by an empirical trajectory analysis of AI systems and capabilities. Ultimately, any framework for effectively regulating AI needs to keep pace with the technology itself.
>
> To those who say there's no problem here to solve, or that California does not have a role in regulating potential national security implications of this technology, I disagree. A California-only approach may well be warranted – especially absent federal action by Congress – but it must be based on empirical evidence and science. The U.S. AI Safety Institute, under the National Institute of Science and Technology, is developing guidance on national security risks, informed by evidence-based approaches, to guard against demonstrable risks to public safety. Under an Executive Order I issued in September 2023, agencies within my Administration are performing risk analyses of the potential threats and vulnerabilities to California's critical infrastructure using AI. These are just a few examples of the many endeavors underway, led by experts, to inform policymakers on AI risk management practices that are rooted in science and fact. [. . .]

Working Group Draft Report on AI Frontier Models. Following his veto of SB 1047, Governor Newsom commissioned the Joint California Policy Working Group on AI Frontier Models to prepare a report on the regulation of frontier models. The Working Group is led by Dr. Fei-Fei Li, Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence; Dr. Mariano-Florentino Cuéllar, President of the Carnegie Endowment for International Peace; and Dr. Jennifer Tour Chayes, Dean of the UC Berkeley College of Computing, Data Science, and Society. In March 2025, the Working Group released a draft of the report and solicited public feedback.[90] Among other things, the draft emphasizes the need for an evidence-based approach to the regulation of frontier models and the importance of balancing regulation and innovation. To help overcome the "evidence dilemma," the draft report calls for enhanced transparency through measures such as whistleblower protections, third-party evaluations, public-facing information sharing, and adverse-event reporting. These measures are addressed in more detail in the following section. The final report is expected in June 2025.

---

[90] "Draft Report of the Joint California Policy Working Group on AI Frontier Models" (Mar. 18, 2025), https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf. ("Draft Report.")

Backlash to AI regulation. The reelection of President Donald Trump heralded an abrupt shift away from efforts to regulate AI. Upon reassuming office, President Trump immediately dismantled his predecessor's efforts to regulate AI, rescinding President Biden's EO and issuing his own Executive Order, "Removing Barriers to American Leadership in Artificial Intelligence."[91]

This shift was evident at the 2025 AI Action Summit in Paris – the third in a series of summits that had previously focused on AI safety and led to the publication of the International AI Safety Report. Tech columnist Kevin Roose reported, "the doomers have been sidelined in favor of a sunnier, more optimistic vision of the technology's potential." According to Roose:

> Panelists and speakers were invited to talk up A.I.'s ability to accelerate progress in areas like medicine and climate science, and gloomier talks about A.I. takeover risks were mostly relegated to unofficial side events. And a leaked draft of the official summit statement, which was expected to be signed by some of the attending nations, was panned by A.I. safety groups for paying too little attention to catastrophic risks.[92]

At the summit, Vice President J.D. Vance criticized the EU AI Act and declared, "The AI future is not going to be won by hand-wringing about safety."[93] French President Emanuel Macron announced plans to invest more than 100 billion Euros ($104 billion) into France's AI sector, declaring "We are committed to go faster and faster."[94]

Here in the US, the de-regulatory agenda is the subject of recently proposed legislation: Congress is currently considering a 10-year moratorium on all state regulation of AI generally.[95]

C. Specific types of risk mitigations for frontier models

Transparency Requirements. Stating that "[t]ransparency is a fundamental prerequisite of social responsibility and accountability,"[96] the Working Group's Draft Report recommends transparency requirements for five categories of information: (1) data acquisition, (2) safety practices, (3) security practices, (4) pre-deployment testing by developers and third parties, and (5) downstream impacts, including disclosures from entities that host foundation models for download or use.[97] Anthropic, a leading AI safety lab, writes:

> In line with the report's findings, we believe governments could play a constructive role in improving transparency in the safety and security practices of frontier AI companies. At

---

[91] Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence" Federal Register (Jan. 23, 2025), https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence.
[92] Kevin Roose, "5 Notes From the Big A.I. Summit in Paris," *The New York Times* (Feb. 10, 2025), https://www.nytimes.com/2025/02/10/technology/ai-summit-paris-technology.html.
[93] The American Presidency Project, "Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France" (Feb. 11, 2025) https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france.
[94] https://observer.com/2025/02/paris-ai-opportunities-overtake-risks/.
[95] Shira Stein & Sophia Bollag, "California state lawmakers ask Congress not to ban their AI laws," *San Francisco Chronicle* (May 21, 2025), https://www.sfchronicle.com/politics/article/bi-partisan-group-state-lawmakers-urges-congress-20335418.php.
[96] Draft Report, *supra*, at p. 25.
[97] *Id.* at pp. 21-23.

> present frontier AI companies are not required to have a safety and security policy (even one entirely of their choice), nor to describe it publicly, nor to publicly document the tests they run – and therefore not all companies do. We believe this could be done in a light-touch way that does not impede innovation. As we wrote in our recent policy submission to the White House, we believe powerful AI systems will arrive soon - perhaps as early as the end of 2026 – so it is important we all devote effort to building a policy regime that creates greater transparency about the safety and security protocols of how AI systems are built.[98]

Whistleblower Protections. The Draft Report also addresses the need for whistleblower protections for employees and contractors of foundation model developers. Because some actions may clearly pose a risk that violates internal company policies but not existing law, the report advises policymakers to "consider protections that cover a broader range of activities."[99]

Adverse Event Reporting Requirements. The Draft Report finds that adverse event reporting – "a proactive monitoring system designed to collect information about relevant events or incidents from various mandated or voluntary reporters"[100] – could improve identification of harms, encourage proactive mitigation, improve coordination between the government and private sector in mitigating risks, and would impose limited costs on reporting entities and the government.[101]

Third-party assessments. The Draft Report states: "For a nascent and complex technology being developed and adopted at a remarkably swift pace, developers alone are simply inadequate at fully understanding the technology and, especially, its risks and harms."[102] The report identifies three strengths to third party evaluation:

> First, third-party evaluations have unmatched scale: Thousands of individuals are willing to engage in risk evaluation, dwarfing the scale of internal or contracted teams. Second, third-party evaluations have unmatched diversity, especially when developers primarily reflect certain demographics and geographies that are often very different from those most adversely impacted by AI. Broad demographic, institutional, and disciplinary diversity is vital for unearthing blind spots. And finally, third-party evaluation is distinctively independent: Society requires forthright and trustworthy assessments of risk, which benefits from a lack of commercial and contractual entanglement with AI developers.[103]

Third-party evaluations benefit the public and foundation model developers: "By establishing industry-wide transparency standards and third-party verification mechanisms, companies can demonstrate compliance with best practices, potentially reducing their liability exposure compared to the uncertainties of purely reactive litigation."[104] Finally, "[t]his transparency coupled with third-

---

[98] Anthropic, "Anthropic's Response to Governor Newsom's AI Working Group Draft Report" (Mar 19, 2025), https://www.anthropic.com/news/anthropic-s-response-to-governor-newsom-s-ai-working-group-draft-report
[99] Draft Report, *supra*, at p. 25.
[100] *Id.* at p. 27.
[101] *Id.* at pp. 28-29.
[102] *Id.* at p. 23.
[103] *Ibid.*
[104] *Ibid.*

party verification effectively creates a 'race to the top' rather than a 'race to the bottom' in safety practices, benefiting responsible companies while improving overall industry standards."[105]

In this regard, civil society, industry, and governmental groups all have expressed support for independent third-party audits of AI systems to ensure that the products can be deployed safely and reliably. In 2024, a KPMG survey of over 1,800 companies across ten major markets found that 91% of business leaders believe that regular audits are the most effective practice in ensuring ethical AI use. Moreover, 80% of business leaders believe that third-party review will be an integral part of that practice.[106] In fact, many AI companies use, and advocate for, independent evaluation.[107] According to Anthropic:

> [W]e believe that [internal self-assessment] is insufficient as it relies on self-governance decisions made by single, private sector actors. Ultimately, testing will need to be done in a way which is broadly trusted, and it will need to be applied to everyone developing frontier systems. This type of industry-wide testing approach isn't unusual - most important sectors of the economy are regulated via product safety standards and testing regimes, including food, medicine, automobiles, and aerospace.[108]

As former OpenAI board members Helen Toner and Tash McCauley have written, "based on experience, we believe that self-governance cannot reliably withstand the pressure of profit incentives."[109] Rather than having AI developers "grade their own homework," supporting the maturation of a third-party auditor ecosystem can play a key role towards requiring robust assessments of frontier models.[110]

Red-teaming. A key tool for identifying risk is red-teaming, a form of adversarial testing that attempts to identify and exploit system vulnerabilities. Successful red-teaming relies on having both internal and external parties trying to "break" the AI system. A recent white paper from University of California Berkeley Center for Long-Term Cybersecurity detailed the importance of red-teaming for frontier models to identify backdoors in systems that could potentially aid in the production of chemical, biological, radiological, nuclear weapons or cyberattacks.[111] Specifically, the paper highlights that red-teaming can bring in experts to perform intensive and interactive testing to better understand the behaviors of the model and offer guidance on mitigating possible existential threats. Similarly, OpenAI recently published a report detailing the importance of third-party red-teaming

---

[105] *Ibid.*
[106] KPMG, "Navigating The AI Era In Financial Reporting," (May 8, 2024), https://kpmg.com/us/en/media/news/ai-in-financial-reporting-kpmg-2024.html.
[107] OpenAI, "OpenAI Red Teaming Network," (Sept. 23, 2023), https://openai.com/index/red-teaming-network/; Anthropic, "Third-party testing as a key ingredient of AI policy." (Mar. 25, 2024), https://www.anthropic.com/news/third-party-testing.
[108] Anthropic, "Third-party testing as a key ingredient of AI policy." (Mar. 25, 2024), https://www.anthropic.com/news/third-party-testing.
[109] Helen Toner & Tasha McCauley, "AI firms mustn't govern themselves, say ex-members of OpenAI's board," *The Economist* (May 26, 2024), https://www.economist.com/by-invitation/2024/05/26/ai-firms-mustnt-govern-themselves-say-ex-members-of-openais-board
[110] https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/.
[111] Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, Jessica Newman, "Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models", *arXiv* (May 15, 2024), https://doi.org/10.48550/arXiv.2405.10986.

for building public trust and collaborating with domain experts who can help surface hard-to-detect risks.[112]

Despite its advantages, red-teaming can be resource-intensive, requiring specialized expertise and substantial time investment. Even then, some vulnerabilities may go undetected. To address this challenge, there have recently been major strides made in development of AI tools to aid in red-teaming. Researchers at MIT have developed a machine learning algorithm capable of generating vast numbers of adversarial prompts designed to break chatbots.[113] This algorithm operates based on "curiosity," when it identifies a prompt that elicits a specific response, it seeks additional prompts that produce similar results. This could be used to identify ways to circumvent built-in guardrails in AI models and provide developers with massive amounts of information regarding the reliability of their systems, data that would be difficult to obtain through human red teaming alone. Together, manual and automated red-teaming can work together to ensure that risks associated with frontier models are properly assessed and mitigated before deployment.

[112] Lama Ahmad, Sandhini Agarwal, Michael Lampe, Pamela Mishkin, "OpenAI's Approach to External Red Teaming for AI Models and Systems," *arXiv* (Jan. 24, 2025), https://doi.org/10.48550/arXiv.2503.16431.
[113] Adam Zewe, "A faster, better way to prevent an AI chatbot from giving toxic responses", *MIT News* (Apr. 10, 2024) https://news.mit.edu/2024/faster-better-way-preventing-ai-chatbot-toxic-responses-0410.