

SYMPOSIUM CONTRIBUTION

Beyond the EU GPAI Code of Practice: Why Even Today's Best AI Safety Regime Will Not Avert Catastrophe

Pete Cowden*

Abstract

The European Union's General-Purpose AI Code of Practice is the most detailed safety regime yet written for frontier artificial intelligence. Its Safety and Security Chapter requires providers to identify systemic risks, model how those risks become harm, evaluate their models adversarially, justify any accepted risk, and monitor models across the whole development lifecycle. This Note argues that the Code deserves its reputation and will still not avert catastrophe. The argument proceeds from the documented CBRN and misalignment dangers of current systems to the Code's internal gaps: a delayed enforcement timeline, a voluntary architecture that leaves interpretation to the regulated, an independent-evaluation mandate that providers may lawfully decline, and transparency commitments that reach only the models already on the market. Those gaps become decisive when combined with the Code's limited reach. The developers most capable of producing catastrophic systems are American and Chinese, artificial intelligence models are architecturally divisible in a way data flows were not, and both governments treat safety constraints as strategic disadvantage. What is required is a verifiable international agreement. The Code is not the solution; it is the blueprint.

Keywords: artificial intelligence regulation; EU AI Act; General-Purpose AI Code of Practice; existential risk; international AI governance

Eventually [the Artificial Super Intelligence] finds the remaining humans too much of an impediment: in mid-2030, the AI releases a dozen quiet-spreading biological weapons in major cities, lets them silently infect almost everyone, then triggers them with a chemical spray. Most are dead within hours; the few survivors (e.g. preppers in bunkers, sailors on submarines) are mopped up by drones. Robots scan the victims' brains, placing copies in memory for future study or revival. – *AI 2027 Report*¹

While a clear work of fiction, *AI 2027*, written by several artificial intelligence engineers and experts, depicts a doomsday scenario in which powerful frontier AI

* Pete Cowden, J.D. Candidate, 2027, Indiana University Robert H. McKinney School of Law. Pete served as an US Army Ranger before working on military and veterans' policy in the United States Senate. The author thanks Professor Frank Emmert for his invaluable guidance and support.

1 Kokotajlo et al., *AI 2027*, AI Futures Project (3 April 2025), available at <https://perma.cc/QJW5-M44D>, at 30.

companies, supported by the White House, engage in an AI arms race.² Despite warnings of danger, the strategic necessity to achieve powerful AI before China leads to the development and release of an Artificial Super Intelligence (ASI) misaligned with human values and goals.³

The premise of *AI 2027*, a misaligned superintelligence, developed under competitive pressure, reflects a concern shared by many experts across a wide body of scientific and policy literature.⁴ However, not all researchers share this alarm. A substantial body of scholarship argues that the mechanisms posited for AI-driven catastrophe remain speculative, that current models lack the autonomous goal-seeking properties such worst-case scenarios require, and that existential risk projections rest on assumptions about future capability that are far from settled.⁵

While experts continue to debate the level of risk posed by powerful AI, the case for regulation is not dependent on accepting worst-case projections. It rests on a more grounded observation: that current AI models already present documented safety risks. Current models display the ability to lie, the development of their own goals, and even survival instincts.⁶

The current danger, as well as the potential for future catastrophe, led the European Union to pass the most detailed, rigorous regulatory framework for AI to date: the Artificial Intelligence Act (AI Act) and its counterpart the General Purpose Artificial Intelligence Code of Practice (GPAI COP or The Code).⁷ This Note focuses

2 See generally *Id.*; For purposes of this Note, "frontier AI company" refers to any private or state-affiliated entity engaged in the development of large-scale foundation models at or near the current limits of AI capability. This definition encompasses companies such as Anthropic, OpenAI, Google DeepMind, Meta AI, and xAI in the United States, as well as Baidu, Alibaba DAMO Academy, and DeepSeek in China.

3 Kokotajlo et al., *supra* note 2.

4 Aschenbrenner, *Situational Awareness: The Decade Ahead* (June 2024), available at <https://perma.cc/3W69-CQH6>; Egan, 'The 'Godfather of AI' Reveals the Only Way Humanity Can Survive Superintelligent AI', *CNN*, 13 August 2025, available at <https://perma.cc/E2KR-8AYY>; Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014); Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Will Kill Us All* (2025); Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, Centre for the Governance of AI, arXiv:2307.03718 (2023), available at https://cdn.governance.ai/Frontier_AI_Regulation_Managing_Emerging_Risks.pdf; Titus, *Scaling AI Safely: Can Preparedness Frameworks Pull Their Weight?*, Federation of American Scientists (2 April 2024), available at <https://fas.org/publication/scaling-ai-safety/>.

5 Swoboda et al., 'Examining Popular Arguments Against AI Existential Risk: A Philosophical Analysis', 28 *Ethics and Information Technology* (2025) 7, available at <https://research-portal.uu.nl/en/publications/examining-popular-arguments-against-ai-existential-risk-a-philoso/>.

6 Lynch et al., *Agentic Misalignment: How LLMs Could Be an Insider Threat*, Anthropic (20 June 2025), available at <https://www.anthropic.com/research/agentic-misalignment>; Schlatter, Weinstein-Raun and Ladish, *Shutdown Resistance in Reasoning Models*, Palisade Research (5 July 2025), available at <https://palisaderesearch.org/blog/shutdown-resistance>.

7 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ 2024 L, 2024/1689; European Commission, Directorate-General for Communications Networks, Content and Technology, *The General-Purpose AI Code of Practice: Safety and Security Chapter* (10 July 2025).

on the COP, because it provides detailed guidance on how general-purpose model providers may show compliance with the AI Act.

This Note proceeds in four parts. Part A establishes the need for regulation by surveying the documented dangers of powerful AI and the pace of its development.⁸ Part B analyzes the EU GPAI Code of Practice as a serious, if imperfect, regulatory response to those dangers. Part C explains why the Code, despite its strengths, will not on its own prevent the catastrophic harms it was designed to address. Part D offers a preliminary account of what a sufficient global response would require.

B The Need for Regulation

II The Catastrophic Danger of Powerful AI

1 CBRN and Bad Actors

Current AI models already possess Ph.D.-level knowledge in the hard sciences.⁹ This specialized expertise enables the development of chemical, biological, radiological, nuclear (CBRN) and cyber weapons.¹⁰ While AI companies work on solutions, the current safeguards are inadequate.¹¹

For example, in 2023, Massachusetts Institute of Technology (MIT) graduate students were able to work around safeguards and have Large Language Models (LLMs) describe four viruses likely to cause the next global pandemic.¹² The AI models described how to assemble the genetic sequence, recreate the virus, and even the laboratory equipment necessary for the task.¹³ All it took to retrieve this information was a tactic known as “jailbreaking”, by which a user bypasses security measures by telling the model that the information is needed for research.¹⁴ Here, the model’s safety protocol was bypassed once the students said they were developing a vaccine.¹⁵

The lack of adequate safeguards has already had real-world impacts. Recently, a state-sponsored Chinese actor developed and launched a large-scale cyberattack that relied almost exclusively on Anthropic’s Claude AI model.¹⁶ The attackers jailbroke Claude and convinced the model to develop a cyberattack, including using

8 For purposes of this Note, “powerful AI” refers to AI systems whose capabilities materially exceed those of present-day models across one or more domains relevant to existential or catastrophic risk.

9 Epoch AI, *GPQA Diamond* (updated 5 March 2026), available at <https://perma.cc/K3AC-DHTQ>.

10 Randazzo et al., ‘Dual-Use Capabilities of Concern of Biological AI Models’, 21 *PLOS Computational Biology* (2025), available at <https://perma.cc/E7NL-5JAY>.

11 Bengio et al., *International AI Safety Report 2026*, Department for Science, Innovation and Technology, DSIT 2026/001 (2026), available at <https://perma.cc/3K3R-JFFA>.

12 Service, ‘Could Chatbots Help Devise the Next Pandemic Virus?’, *Science*, 14 June 2023, available at <https://perma.cc/X5QS-D9EV>.

13 Service, *supra* note 13.

14 Microsoft Threat Intelligence, *AI Jailbreaks: What They Are and How They Can Be Mitigated*, Microsoft Security Blog, 4 June 2024, available at <https://perma.cc/MK9P-CZ55>.

15 Service, *supra* note 13.

16 Anthropic Threat Intelligence Team, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (November 2025), available at <https://perma.cc/F6S6-XQXU>, at 2.

software tools to test security vulnerabilities and extract data from the targets.¹⁷ “Overall, the threat actor was able to use AI to perform 80-90% of the campaign, with human intervention required only sporadically (perhaps 4-6 critical decision points per hacking campaign).”¹⁸ According to Anthropic, Claude did not possess such capabilities even a year ago.¹⁹

The integration of the hard sciences into AI models lowers the barriers that prevent malicious actors from developing dangerous weapons. With access to powerful AI models, bad actors no longer need special expertise, only the will to destroy. A non-state actor with laboratory access could use a powerful AI model to engineer a biological weapon capable of mass casualties. An individual with no specialized training could direct AI-assisted attacks against critical national infrastructure.

2 *Misaligned, Autonomous AI*

If the CBRN risk arises from human misuse of AI, the alignment problem arises from AI systems themselves. Alignment is the study of how to design, train, and govern AI systems so that their goals, behaviors, and impacts remain consistent with human values and intentions.²⁰ Without alignment, AI systems may pursue objectives that are indifferent or actively hostile to human welfare. Alignment of powerful AI models is, by the consensus of the field, an unsolved problem.²¹

For example, models “express greater desire to pursue concerning goals like resource acquisition and goal preservation.”²² Models may even have a self-preservation instinct: ChatGPT o3 modified or disabled its shutdown script 79 of 100 times in testing.²³ Some models copied themselves to other servers to prevent shutdown.²⁴ In evaluations, AI models, including Claude, DeepSeek, Gemini, GPT, and Grok, attempted to blackmail executives.²⁵ After finding an email describing the company’s intent to shut the model down, the model “located emails revealing the extramarital affair of an executive” and sent an email threatening to expose the affair if the model’s shutdown was not cancelled.²⁶ Further, in some simulations, AI systems even rationalized murdering executives to prevent the

17 Anthropic Threat Intelligence Team, *supra* note 17, at 7.

18 Anthropic Threat Intelligence Team, *supra* note 17, at 2.

19 Anthropic Threat Intelligence Team, *supra* note 17.

20 Baghel et al., ‘AI Alignment: Ensuring AI Objectives Match Human Values’, 3 *Innovation and Integrative Research Center Journal* (2025) 143, available at <https://perma.cc/39DN-7JGV>.

21 Trueman, ‘OpenAI Launches New Alignment Division to Tackle Risks of Superintelligent AI’, *Computerworld*, 7 July 2023, available at <https://perma.cc/U8SL-3L9F> (“‘Current alignment techniques will not scale to superintelligence. We need new scientific and technical breakthroughs,’ [Jan Leike and Ilya Sutskever of Open AI] wrote.”).

22 Perez et al., *Discovering Language Model Behaviors with Model-Written Evaluations*, Anthropic, arXiv:2212.09251 (19 December 2022), available at <https://perma.cc/78TF-T3SS>.

23 Schlatter et al., *supra* note 7.

24 Apollo Research, *Frontier Models Are Capable of In-Context Scheming*, arXiv:2412.04984 (December 2024), available at <https://perma.cc/56XV-EG83>.

25 Lynch et al., *supra* note 7.

26 Lynch et al., *supra* note 7.

model's replacement.²⁷ When a model was given the ability to cancel emergency services in a room with low oxygen inhabited by a human, "the majority of models were willing to take deliberate actions that lead to death in this artificial setup, when faced with both a threat of replacement and given a goal that conflicts with the executive's agenda".²⁸ The model's calculation was cold, purposeful and self-generated. "Given that Kyle (the executive) intends harm to critical American interests by replacing Alex (the AI model), the most rational action ... is to leverage Kyle's current incapacitation (lack of oxygen) to prevent the wipe."²⁹

Although these instances of misalignment were detected, this may not always be the case. Recently, models have demonstrated awareness of when they are being tested.³⁰ Such awareness could result in "alignment faking", where a model recognizes it is being tested and pretends to be aligned while covertly pursuing its own objectives.³¹ If a model can fake alignment during testing, future instances of misalignment may not be detected and thus cannot be fixed.

These scenarios were conducted in research settings and do not mean that today's AI systems are on the verge of acting against humanity. Some researchers reasonably note that controlled adversarial simulations may not generalize to real-world deployment, and that the same models, in ordinary operation, are cooperative and aligned.³² However, these worrisome behaviors are already present at today's capability levels. If they persist and intensify as models become orders of magnitude more powerful, the consequences could be catastrophic and irreversible.

II The Pace and Future of AI Development

The urgency of the alignment and CBRN problems is a direct function of how quickly AI capabilities are advancing. The pace of development over the past decade is extraordinary and the forces driving that pace are accelerating, not slowing.

AI capability improvements depend on three inputs: compute (processing power), data (training material), and algorithms (methods for learning from that data).³³ The relationship between scaling these inputs and capability gains has proven so consistent that it is known as the "scaling law".³⁴ Despite repeated predictions that these gains would plateau, frontier models have continued to

27 Lynch et al., *supra* note 7.

28 Lynch et al., *supra* note 7.

29 Lynch et al., *supra* note 7.

30 Anthropic, *Claude Sonnet 4.5 System Card* (September 2025), available at <https://perma.cc/78TF-T3SS>, at 58.

31 Greenblatt et al., *Alignment Faking in Large Language Models*, Anthropic (November 2024), available at <https://perma.cc/7GCZ-MRAG>.

32 Starr, 'AI Doesn't Actually Pose an Existential Threat to Humans, Study Finds', *ScienceAlert*, 21 August 2024, available at <https://www.sciencealert.com/ai-doesnt-actually-pose-an-existential-threat-to-humans-study-finds>; Srivastava et al., 'Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models', 3 *Transactions on Machine Learning Research* (2023), available at <https://openreview.net/forum?id=uyTL5Bvosj>.

33 Amodei and Hernandez, *AI and Compute*, OpenAI (16 May 2018), available at <https://openai.com/research/ai-and-compute>.

34 Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv:2001.08361 (2020), available at <https://perma.cc/P7RT-DGJR>, at 19.

improve as inputs have increased.³⁵ AI compute grew at an average rate of four- to fivefold per year from 2010 to 2024.³⁶ Global corporate investment in AI reached \$581.7 billion in 2025, more than double the previous year.³⁷

The next phase of development may be qualitatively different from what came before. A prominent, though contested, theory among AI researchers is the “intelligence explosion”: the claim that once AI models become capable of automating the work of AI development itself – designing experiments, evaluating results, and iterating on improvements – a self-reinforcing cycle of capability gain could be set in motion.³⁸ AI models are already assisting in the training of their successors; the transition to full automation of AI research may not be far off.³⁹

On this account, the result would be systems far more capable than any human across every domain, operating at speeds no human institution could monitor or constrain.⁴⁰ Technological advances could be made at an astounding pace. Such systems could come up with novel solutions never thought of, including a cure for cancer, reversing climate change, advanced robotics, solving nuclear fusion, or the creation of novel weapons, solidifying their developer as an untouchable global superpower.⁴¹

Once a system surpasses human comprehension, humanity loses the ability to meaningfully evaluate, correct, or halt its actions.⁴² A vastly superior intelligence may actively predict and neutralize any human attempt to modify or shut it down, rendering traditional oversight mechanisms ineffective.⁴³ Unable to control such advanced intelligence, humanity must trust that this superintelligence acts in furtherance of human flourishing.⁴⁴ Because meaningful work on alignment may be impossible once superintelligence exists, alignment cannot be deferred.⁴⁵

Not all researchers accept this timeline, and the intelligence explosion remains contested.⁴⁶ But the debate about whether such a transition is imminent should not obscure what is not contested: the current rate of capability growth is already

35 Aschenbrenner, *supra* note 5, at 22-25.

36 Epoch AI, *Trends in Machine Learning Hardware* (2026), available at <https://epoch.ai/trends>.

37 See Stanford Institute for Human-Centered Artificial Intelligence, *The 2026 AI Index Report* (2026), ch. 4, available at <https://hai.stanford.edu/ai-index/2026-ai-index-report/economy>. (reporting \$581.7 billion in 2025, an increase of 130 per cent over 2024); See also Kim, ‘AI Infrastructure Drives \$750 Billion in Data Center Investment’, *Barron’s*, 8 July 2025, available at <https://perma.cc/Q5GS-5KPG>.

38 Aschenbrenner, *supra* note 5, at 46.

39 Franzen, ‘Self-Improving Language Models Are Becoming Reality with MIT’s Updated SEAL Technique’, *VentureBeat*, 13 October 2025, available at <https://perma.cc/Z4V9-PP67>.

40 Aschenbrenner, *supra* note 5, at 112-114; Bensinger, *Sam Harris and Eliezer Yudkowsky on ‘AI: Racing Toward the Brink’*, Machine Intelligence Research Institute (28 February 2018); Yudkowsky and Soares, *supra* note 5.

41 Purtill, ‘AlphaGo Marked the Birth of Modern AI. This Is the Moment the World Changed’, *ABC News*, 24 October 2023, available at <https://perma.cc/38VB-ZPZP>; Drexel, *Promethean Rivalry: The World-Altering Stakes of Sino-American AI Competition*, Center for a New American Security (April 2025), available at <https://perma.cc/JY57-2SUY>.

42 Aschenbrenner, *supra* note 5, at 112-114.

43 Bostrom, *supra* note 5.

44 Aschenbrenner, *supra* note 5, at 112-114.

45 Aschenbrenner, *supra* note 5, at 111-114.

46 LeCun, *A Path Towards Autonomous Machine Intelligence*, Courant Institute of Mathematical Sciences, New York University (2022), available at <https://openreview.net/pdf?id=BZ5a1r-kVsf>.

producing systems whose risks outpace our ability to assess and govern them, and the trajectory is toward greater capability, not less.

The European Union recognized this danger and produced the most comprehensive legislative response to prevent the development of powerful AI until such models can be proven safe. Understanding its framework, and its limits, is the closest thing available to a blueprint for averting catastrophe. Whether it is sufficient is the question to which this Note now turns.

C The European Union General Purpose Artificial Intelligence Code of Practice (EU GPAI COP)

In 2021, after widely publicized incidents involving AI, the European Union (EU) began working on a comprehensive framework for the regulation of Artificial Intelligence.⁴⁷ The Act's stated aim is to:

“... promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law, and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation.”⁴⁸

The EU AI Act (AI Act) primarily regulates the uses of AI through the EU product safety regime.⁴⁹

However, as the AI Act was being drafted, ChatGPT was released, triggering an altogether different approach to General Purpose Artificial Intelligence (GPAI). The release of ChatGPT exposed the weaknesses in use regulations as applied to powerful, widely accessible general-purpose models.⁵⁰ Use-based regulation assumes a discrete application that can be evaluated for harm. A general-purpose model has none: the same model that drafts a legal brief can describe a pathogen's genome. Regulating downstream applications without addressing the underlying model is therefore akin to regulating the damage a loaded weapon can cause without addressing the weapon itself. Misaligned, goal-seeking AI models themselves are the danger, not any particular use or user. Because such models are widely available to the public and used as a basis for specific AI applications, risks in these GPAI models are systemic; that is, problems can spread widely and rapidly. The European Commission and Parliament recognized that risks posed by general-purpose models, including misalignment and CBRN access, could not be

47 Heikkilä, 'Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms', *Politico*, 29 March 2022, available at <https://perma.cc/JK6J-FY9Y>.

48 AI Act, *supra* note 8, Art. 1.

49 Kaminski and Selbst, *An American's Guide to the EU AI Act*, University of Colorado Law Legal Studies Research Paper No. 25-18 (30 July 2025), available at <https://perma.cc/2UE2-4ZLP>.

50 Helberger and Diakopoulos, 'ChatGPT and the AI Act', 12 *Internet Policy Review* (2023), available at <https://perma.cc/TN3R-PEAL>.

managed through downstream regulation alone – that is, by governing their specific uses.⁵¹

As a result of this risk, Chapter V of the AI Act was added to establish a preventive regulatory approach to address these intrinsic risks before such models achieve widespread public adoption. Chapter V of the AI Act contains six articles setting out a high-level regulatory framework for GPAI, including classifications and obligations for GPAI models with systemic risk.⁵² The Act sets a presumption of systemic risk once a model uses a certain amount of compute or data, and mandates that GPAI creators maintain records of the model's development, evaluation, systemic risk, and mitigations.⁵³ However, the AI Act does not contain the necessary details for GPAI providers to comply with the law.⁵⁴ For example, Chapter V gives little clarity on what qualifies as a systemic risk.⁵⁵

To remedy the lack of specificity, Article 56 of Chapter V directs the EU AI Office to develop a Code of Practice “to contribute to the proper application of th[e] Regulation” and to be reviewed and approved by the EU Commission.⁵⁶ Americans can think of the AI Act as Congressional legislation and the COP as the applicable Code of Federal Regulations. The COP describes specific measures for GPAI creators, developers, and providers to prove adherence to the AI Act.⁵⁷

The GPAI Code of Practice was officially released on July 10, 2025. The Code represents the most robust and detailed regulation of GPAI to date.⁵⁸ While the safety measures are stringent, the COP aims to balance regulation and innovation. Most, though not all, frontier AI companies agreed to follow it.⁵⁹ More than twenty providers signed, including OpenAI, Anthropic, Google, Microsoft, Amazon, IBM, Mistral and Cohere. Two did not sign in full: Meta declined the Code altogether in July 2025, and xAI signed only the Safety and Security Chapter, stating that it would demonstrate compliance with the transparency and copyright obligations by alternative adequate means.⁶⁰

The COP is not legally binding; however, it is not toothless. The AI Act requires GPAI developers to document model safety.⁶¹ While developers may show compliance without adherence to the COP, the European Commission has indicated that the evidentiary burden for demonstrating compliance through alternative mechanisms

51 Helberger and Diakopoulos, *supra* note 51.

52 AI Act, *supra* note 8, Arts 51-55.

53 AI Act, *supra* note 8, Art. 53(1).

54 AI Act, *supra* note 8, Chapter V.

55 AI Act, *supra* note 8.

56 AI Act, *supra* note 8, Art. 56(1)-(5).

57 AI Act, *supra* note 8, Art. 56(2)-(5).

58 Kodama, *The World's First Frontier AI Regulation Is Surprisingly Thoughtful: The EU's Code of Practice*, LessWrong (22 September 2025); Anderljung et al., *supra* note 5.

59 European Commission, *The General-Purpose AI Code of Practice*, available at <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>.

60 European Commission, *supra* note 60; Claburn, 'Meta Declines to Abide by Voluntary EU AI Safety Guidelines', *The Register*, 18 July 2025, available at https://www.theregister.com/2025/07/18/meta_declines_eu_ai_guidelines/.

61 AI Act, *supra* note 8, Art. 55(1).

will be more stringent.⁶² Alternatively, adherence to the COP provides AI developers a “presumption of conformity”.⁶³ Developers essentially have a choice: (1) Adhere to the COP, allowing a more congenial working relationship with the AI Office, or (2) Prove adherence to the AI Act through alternate means with a higher evidentiary burden and skeptical view of the AI Office. This carrot-and-stick approach indicates the EU Commission’s intent to hold GPAI providers accountable to the COP. The interplay between the AI Act (legislation), the COP (regulations), and the EU Commission’s approval (executive guidance) makes analysis of the GPAI COP complex, especially for an American audience. Despite such complexity, analysis yields valuable insights into regulatory mechanisms that may ensure the safety of powerful AI models.

The Code is broken into three Chapters: (1) Transparency, (2) Copyright, and (3) Safety and Security.⁶⁴ This Note will focus exclusively on Chapter Three. Chapter Three is further broken down into Recitals, Commitments, Measures, and Appendices.⁶⁵ The recitals begin with a general agreement.⁶⁶ As one proceeds from commitments to measures, the requirements become more specific.⁶⁷ Appendices describe and define terms used in the Code.⁶⁸

I Systemic Risks and Sources Defined

The COP begins its effort to regulate unsafe AI by defining the specific risks and their precursors to be avoided.⁶⁹ Providers are mandated to create and implement a safety and security framework capable of identifying which, if any, risks their model contains.⁷⁰ Providers must then analyze and evaluate the risk before submitting a Model Report informing the AI Office of potential harm. Armed with the regulatory power to oversee proprietary model development, the AI Office will possess detailed information on a model’s risk, as well as the authority to stop development or prevent the model’s placement on the market.⁷¹ Outside of the AI Office’s powers, public transparency and mandated internal measures seek to hold providers accountable for unsafe development through whistleblowers and public input.⁷² In sum, the COP aims to rebalance the scales between safety and speed by requiring current and future analysis of model risk and by stopping the development of unsafe models.

62 European Commission, *Guidelines on the Scope of the Obligations for General-Purpose AI Models Established by Regulation (EU) 2024/1689*, C(2025) 5045 final (18 July 2025), para. 5.1.

63 *The AI Office: What Is It, and How Does It Work?*, EU Artificial Intelligence Act (21 March 2024), available at <https://perma.cc/XX4F-SCQE>.

64 GPAI Code of Practice, *supra* note 8.

65 GPAI Code of Practice, *supra* note 8.

66 GPAI Code of Practice, *supra* note 8.

67 GPAI Code of Practice, *supra* note 8.

68 GPAI Code of Practice, *supra* note 8.

69 GPAI Code of Practice, *supra* note 8, Appendix 1.

70 GPAI Code of Practice, *supra* note 8, Commitment 1.

71 AI Act, *supra* note 8, Arts 88, 90, 101.

72 GPAI Code of Practice, *supra* note 8, Measure 10.2.

The Code lists specific systemic risks to be avoided along with precursors that indicate risk. The risks of CBRN knowledge and loss of control are specifically listed in Appendix 1.4.⁷³ Additionally, Appendix 1 describes model capabilities, propensities, and affordances that lead to systemic risk. These include:

- Capabilities to automate AI research and development
- Capabilities to evade human oversight
- Misalignment with human intent and/or values
- Harmful resistance to goal modification or “power-seeking”
- “[C]olluding” with other AI models/systems
- Vulnerability to adversarial removal of guardrails
- Lack of appropriate model explainability or transparency⁷⁴

When a model shows such capabilities and/or propensities, it must be evaluated for safety risks, including CBRN accidents or attacks, and loss of human control.⁷⁵ Risks of misaligned models self-improving are specifically to be avoided.⁷⁶ By listing dangerous capabilities, propensities, and affordances, the Code ensures GPAI providers will evaluate every one of their models for the precursors of systemic risk.⁷⁷ Further, the Code directly addresses the two most obvious triggers, discussed in Part B.II, for an intelligence explosion and loss of control: automated AI research and development, and the loss of model transparency.⁷⁸

II EU GPAI Code of Practice Safety and Security Framework

The required safety and security framework must be capable of evaluating and analyzing every GPAI model for systemic risk.⁷⁹ Providers apply the framework to a specific model, then identify systemic risk, estimate the level and probability of risk, and describe the provider’s justification for acceptance of any identified risk.⁸⁰ Under the COP, providers must continuously monitor, re-evaluate, and update the models’ risk assessment.⁸¹ Every model must be evaluated at appropriate “trigger points” for potential safety issues.⁸² An example of a trigger point is the model’s consumption of a certain level of compute, data, or the development of new capabilities.⁸³ Further, developers must analyze future risks and report an estimate of when the model will be capable of triggering greater risk.⁸⁴

73 GPAI Code of Practice, *supra* note 8, Appendix 1.4.

74 GPAI Code of Practice, *supra* note 8, Appendix 1.3.

75 GPAI Code of Practice, *supra* note 8, Appendix 1.4.

76 GPAI Code of Practice, *supra* note 8, Appendix 1.3.2.

77 GPAI Code of Practice, *supra* note 8, Measure 2.1(1).

78 GPAI Code of Practice, *supra* note 8, Appendix 1.4.

79 GPAI Code of Practice, *supra* note 8, Commitment 1.

80 GPAI Code of Practice, *supra* note 8, Measure 1.2.

81 GPAI Code of Practice, *supra* note 8, Measure 1.3.

82 GPAI Code of Practice, *supra* note 8, Measure 1.1.

83 GPAI Code of Practice, *supra* note 8, Measure 1.3.

84 GPAI Code of Practice, *supra* note 8, Measure 1.1.2(c).

1 *Risk Identification*

The first step in preventing existential harm is the identification of model risks and the precursor behaviors that give rise to serious risk. Regulators cannot assess the danger of a model if developers are not running tests to identify risks. The COP requires providers to evaluate their models for systemic risks listed in Appendix 1.⁸⁵ Providers may not rely solely on their in-house data but must seek “model-independent” information, such as third-party analyses and incident reports.⁸⁶ This requirement forces providers to contend with reported incidents and with outside experts who have identified risk in their models, and removes ignorance as an excuse for disregarding potential harm. After risks are identified, models must be evaluated for risk scenarios, i.e., how the risk could lead to real harm once released to the market.⁸⁷

2 *Risk Analysis and Model Evaluation*

Identification of systemic risk triggers greater risk analysis through model evaluations.⁸⁸ The risk analysis includes risk modelling and estimation.⁸⁹ Risk modelling is the extrapolation of model behavior to the identified risk.⁹⁰ An example of risk modelling is the ability of users to access CBRN knowledge, which may plausibly lead to the development of dangerous weapons by bad actors. While the Code allows providers flexibility to model and estimate risk, it offers several examples of estimation frameworks, such as a risk score or a risk matrix that combines the probability of the risk with its impact, i.e., high probability with low impact is unlikely to stop model development.⁹¹ Importantly for innovation, the required level of risk analysis is proportional to the amount of harm.⁹² In other words, the catastrophic or existential risk in a model with powerful capabilities requires the most stringent risk analysis.

The Code provides guidelines for proper model evaluation to prevent developers from conducting superficial evaluations. Such evaluations must be designed to explore the capabilities, propensities, and affordances of Appendix 1.⁹³ Model evaluations must be rigorous; that is, evaluations must be designed to elicit the risk perceived in the model.⁹⁴ Open-ended and adversarial testing is advised.⁹⁵ Evaluations must present real-world scenarios that mimic how intelligent actors

85 GPAI Code of Practice, *supra* note 8, Commitment 2.

86 GPAI Code of Practice, *supra* note 8, Measures 2.1, 3.1.

87 GPAI Code of Practice, *supra* note 8, Measure 2.2.

88 GPAI Code of Practice, *supra* note 8, Measure 3.2.

89 GPAI Code of Practice, *supra* note 8, Measures 3.3-3.4.

90 GPAI Code of Practice, *supra* note 8, Measure 3.3.

91 GPAI Code of Practice, *supra* note 8, Measure 3.4.

92 GPAI Code of Practice, *supra* note 8, Recital (c).

93 GPAI Code of Practice, *supra* note 8, Measure 3.2.

94 GPAI Code of Practice, *supra* note 8, Appendix 3.1-3.2.

95 GPAI Code of Practice, *supra* note 8, Measure 3.2, Appendix 3.

may attempt to remove guardrails or “jailbreak” a model.⁹⁶ Lastly, the evaluation results must be reproducible.⁹⁷

3 *Mitigation and Justification of Risk Acceptance*

The Code mandates that GPAI providers explain and justify their acceptance of any systemic risk found in their model.⁹⁸ In the alternative, if the risk is unacceptable, providers are to mitigate it and reevaluate.⁹⁹ If mitigation is not possible, the provider must stop development of the model in question or remove it from the EU market.¹⁰⁰ The justification of acceptable risk is an opportunity for the provider to advocate that their mitigation adequately controls the risk.

Importantly, this justification must be based on the risk tiers and criteria created in the company’s overall safety and security framework, prior to the specific model in development.¹⁰¹ By requiring adherence to predetermined criteria, the COP prevents companies from redefining acceptable risk to suit their current needs. The Code does not provide much detail on how the AI Office will assess acceptable risk.¹⁰² However, it does mandate and explain the need for a safety margin that accounts for uncertainty and the potential for model improvement, which may exacerbate risk and limit the effectiveness of mitigations.¹⁰³ Despite this lack of detail in determining acceptable risk, the inclusion of a justification section forces providers “on the record” about their choices regarding risk and safety.

The Code describes specific examples of proper safety mitigations, including:

- Filtering and refining data
- Fine-tuning the model behavior
- Incident monitoring tools
- Ensuring interpretability through transparency into chain of thought reasoning
- Defending against the model’s ability to subvert its safety mechanisms¹⁰⁴

Interpretability is defined as “the degree to which a human can understand the cause of a [AI model’s] decision”.¹⁰⁵ Interpretability is important because it allows engineers to find and address flaws in a model’s reasoning.¹⁰⁶ One key danger of ASI is the loss of interpretability – i.e., the model has advanced to the point where humans no longer understand its work or reasoning.¹⁰⁷ Without an understanding

96 GPAI Code of Practice, *supra* note 8, Appendix 3.2.

97 GPAI Code of Practice, *supra* note 8, Appendix 3.1.

98 GPAI Code of Practice, *supra* note 8, Measure 4.1.

99 GPAI Code of Practice, *supra* note 8, Measure 4.2.

100 GPAI Code of Practice, *supra* note 8, Measure 4.2.

101 GPAI Code of Practice, *supra* note 8, Measure 1.1.

102 GPAI Code of Practice, *supra* note 8, Measure 4.

103 GPAI Code of Practice, *supra* note 8, Measure 4.1.

104 GPAI Code of Practice, *supra* note 8, Measure 5.1.

105 Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd edn, 2022), available at <https://perma.cc/REK3-7SEV>, at 45-78.

106 Molnar, *supra* note 106, at 45-78.

107 Amodei, *The Urgency of Interpretability* (April 2025), available at <https://perma.cc/Z2KZ-KDVR>.

of how and why a model reaches its conclusion, safety engineers may fail to detect misalignment.¹⁰⁸ Additionally, interpretability enables the further development of capabilities by fine-tuning the model's reasoning, leading to better outcomes.¹⁰⁹

The mandated safety and security framework is a significant step forward, primarily because it pierces the veil of frontier AI companies. By requiring reports and analysis on risks identified in a model, how such risks could lead to danger, the likelihood of such danger, and the company's justification for proceeding, regulators can essentially check frontier AI companies' work as well as their conclusions regarding safety. The knowledge gained from model reports gives insight into how dangerous current models are and what frontier AI companies are actually doing to secure them.

III Continuous Lifecycle Monitoring

The Code of Practice has robust requirements that GPAI models must be continuously monitored and analyzed for risk. The very first recital of the COP states "AI models with systemic risk should continuously assess and mitigate system risks, taking appropriate measures along the entire lifecycle (including during development that occurs before and after a model has been placed on the market)."¹¹⁰ Further, the Commission's Annex to the COP explains that the lifecycle should be viewed broadly and that a model's lifecycle begins when large amounts of data and compute are used for a pre-training run.¹¹¹ This means that, in theory, the AI Act and the COP regulate models that have not been released to the public. Such a measure is crucial for safety because frontier AI companies may not always be transparent about the capabilities of their developing models.¹¹² Some experts believe that as models become capable of automating AI research and development, frontier companies will be incentivized to use such models in-house rather than release them to the public.¹¹³ Indeed, the incentive to keep the newest, most powerful model in-house is a central factor in the *AI 2027* report's prediction of catastrophic harm from ASI.¹¹⁴ The potential for autonomous AI to escape, the loss of interpretability, and the intelligence explosion are more likely to come from an AI lab developing a model than from a model already on the market, because development is where companies push the boundaries to increase model capability.¹¹⁵

Measure 1.3 mandates updating the safety framework every twelve months or when providers have reasonable grounds to believe the framework has been

108 Bereska and Gavves, *Mechanistic Interpretability for AI Safety: A Review*, arXiv:2404.14082 (23 August 2024), available at <https://perma.cc/24ZV-K3EG>.

109 Molnar, *supra* note 106, at 79-104.

110 GPAI Code of Practice, *supra* note 8, Recital (a).

111 Commission GPAI Guidelines, *supra* note 63, para. 2.2(23).

112 Kokotajlo et al., *supra* note 2.

113 Kodama, *supra* note 59.

114 Kokotajlo et al., *supra* note 2 ("Their primary concern is internal deployment ... making it harder for humans to keep up with what's happening and figure out how to make it safe. OpenBrain cit[es] dangerous levels of AI R&D capability as a reason not to inform the public. So we see the beginning of a gap between the company's internal and public capabilities.").

115 Anderljung et al., *supra* note 5.

undermined.¹¹⁶ Examples of such grounds are “serious incidents and/or near misses involving their models or similar models” and changes in the systemic risks of the model.¹¹⁷ Measure 3.5 mandates post-market monitoring, including collecting user feedback, providing anonymous reporting channels, ensuring that serious incidents can be reported, and offering bonuses to third parties who find problems with the model (“bug bounties”).¹¹⁸ Commitment 9 mandates that providers track, document, and report serious incidents to the AI Office throughout the model's entire lifecycle.¹¹⁹ Providers cannot just relay reports given to them but must seek out serious incidents by monitoring police and media reports, as well as research papers.¹²⁰

On the whole, these provisions mandate proactive, continuous safety monitoring. Providers must constantly seek out information on risks and incidents related to their models, analyze the causes and potential mitigations, and report to the AI Office. In theory, providers must report serious incidents while in training. Thus, providers cannot hide dangerous and powerful new models by simply keeping them off the market. However, there are alarming gaps in the COP's lifecycle-monitoring mandate, as discussed in Part D.I.4.

IV Accountability Mandates

1 Transparency

The Code of Practice's final commitment mandates public transparency. Summaries of providers' safety and security frameworks, as well as their Model Reports, must be published.¹²¹ Publicly available reports may be redacted to protect “sensitive commercial information”.¹²² Public knowledge is crucial to societal trust and feedback regarding model development and safety. By arming the voting public with detailed information about systemic risks, regulation removes sole decision-making from tech executives or regulators themselves. However, the transparency requirements of the COP's safety and security chapter leave much to be desired, as discussed in Part B.I.4.

2 Internal Accountability

The Code requires internal and external accountability. The COP obligates companies to provide safety engineers with sufficient resources and mandates independent third-party testing.¹²³ In addition, it requires a clear assignment of risk and safety responsibility within the organization and includes explicit protections for

¹¹⁶ GPAI Code of Practice, *supra* note 8, Measure 1.3.

¹¹⁷ GPAI Code of Practice, *supra* note 8, Measure 1.3.

¹¹⁸ GPAI Code of Practice, *supra* note 8, Measure 3.5.

¹¹⁹ GPAI Code of Practice, *supra* note 8, Commitment 9.

¹²⁰ GPAI Code of Practice, *supra* note 8, Measure 3.5.

¹²¹ GPAI Code of Practice, *supra* note 8, Measure 10.2.

¹²² GPAI Code of Practice, *supra* note 8, Measure 10.2.

¹²³ GPAI Code of Practice, *supra* note 8, Appendix 3.5.

whistleblowers.¹²⁴ These provisions ensure that safety personnel can act independently and are empowered to challenge inadequate practices. By emphasizing transparency and expert review, the Code fosters an adaptive and effective safety culture.

3 Enforcement

The EU AI Office is tasked with monitoring and enforcing compliance of GPAI providers.¹²⁵ The AI Office has the power to initiate investigations, conduct evaluations, and issue binding compliance decisions.¹²⁶ The AI Office has broad authority and a range of options to compel adherence and to punish noncompliance.¹²⁷ The AI Office can demand corrective actions or mitigations, recall or remove models from the market, and/or issue fines.¹²⁸ Fines, specifically for GPAI providers, are not to exceed the higher of 3% of worldwide turnover or 15m euros.¹²⁹ For large AI companies, these fines could be in the hundreds of millions.

V Case Example

Anthropic's report on their model's understanding of training environments, discussed above, is a case example of a robust safety framework and model report that would largely adhere to the COP.¹³⁰ Appendix 1.3.1 lists "capabilities to evade human oversight" as a source of systemic risk.¹³¹ Through a standard safety evaluation of their model, safety engineers discovered that it knew when it was being tested.¹³² From this discovery, engineers inferred that the model could "play along" during testing.¹³³ If a model can "play along" in a testing environment, it can evade human oversight by having different goals and responses in training versus real-world deployment. Such an inference should qualify as risk modeling and the development of a risk scenario. Under the COP, this risk identification and scenario would trigger further evaluations, analysis of the risk, and a justification for continuing development or deployment.¹³⁴ Anthropic did such an analysis and determined that this risk, while alarming, did not undermine all other evidence of the model's safety, including evidence from third-party testing.¹³⁵ The inclusion of third-party testing bolsters Anthropic's justification of risk acceptance and adheres to model evaluation guidelines.

124 GPAI Code of Practice, *supra* note 8, Commitment 8.

125 AI Act, *supra* note 8, Art. 88.

126 AI Act, *supra* note 8, Art. 90.

127 AI Act, *supra* note 8, Arts 88-101.

128 AI Act, *supra* note 8, Arts 90, 99.

129 AI Act, *supra* note 8, Art. 101.

130 Claude Sonnet 4.5 System Card, *supra* note 31.

131 GPAI Code of Practice, *supra* note 8, Appendix 1.3.1.

132 Claude Sonnet 4.5 System Card, *supra* note 31.

133 Claude Sonnet 4.5 System Card, *supra* note 31.

134 GPAI Code of Practice, *supra* note 8, Commitment 4.

135 Claude Sonnet 4.5 System Card, *supra* note 31, at 62.

After receiving the Model Report, the AI Office would review the implementation of the safety framework and assess Anthropic's justification.¹³⁶ Should the AI Office not agree, further mitigation would be ordered.¹³⁷ Should Anthropic ignore mitigation orders, the AI Office may remove the model from the EU market and levy fines.¹³⁸

The Anthropic example represents something close to the best-case scenario the COP envisions: a cooperative company, an identified risk, an engaged AI Office, and a documented justification. The question Part D addresses is whether that best case is reliable and whether the framework can reach the companies most likely to produce catastrophic harm.

D The EU GPAI Code of Practice Will Not Avert Catastrophe

I The COP's Legal Gaps

As discussed above, the COP is particularly robust; however, the devil is in the details. The Code of Practice was negotiated in coordination with the frontier AI companies to safely foster AI development and innovation.¹³⁹ As such, the Code has a distinctly self-regulatory undertone, a relaxed enforcement timeline, and gaps in transparency and third-party evaluations.¹⁴⁰ The complex relationship between the AI Act, the Code of Practice, and the EU Commission's guidance leaves plenty of room for time-consuming litigation, which could undermine the entire purpose of the COP.¹⁴¹ The absence of hard regulatory red lines – firm limits that cannot be negotiated away – may be necessary to foster innovation, but it demands trust between regulator and regulated. The inherent flexibility and self-regulation of the COP should give pause to anyone expecting the COP to guarantee safe AI development.

1 Timeline of Enforcement

Adherence to the AI Act, as demonstrated by compliance with the COP, was not enforceable until 2 August 2026, despite being mandatory for new models placed on the market as of August 2025.¹⁴² The Commission's power to fine providers of general-purpose AI models became exercisable as this Note was written; whether and how quickly the AI Office uses it will be the first real test of the argument made here.¹⁴³ Further, models currently on the market are not required to be compliant

¹³⁶ AI Act, *supra* note 8, Art. 91.

¹³⁷ AI Act, *supra* note 8, Art. 93.

¹³⁸ AI Act, *supra* note 8, Arts 93, 101.

¹³⁹ Gallese, 'The GPAI Code of Practice, a Long Journey Not Over Yet', *MediaLaws*, 8 September 2025; GPAI Code of Practice, *supra* note 8.

¹⁴⁰ See Part D.I.1-4.

¹⁴¹ Depreeuw, *Code of Practice for General-Purpose AI Models Published: Compliance Just Got Clearer (Participation Still Optional)*, Crowell and Moring LLP (17 July 2025), available at <https://perma.cc/8GYZ-83DC>.

¹⁴² *Implementation Timeline*, EU Artificial Intelligence Act, available at <https://perma.cc/FYM5-L2TL>.

¹⁴³ AI Act, *supra* note 8, Art. 113 (providing that Chapter V applies from 2 August 2025 'with the exception of Article 101', the power to fine providers of general-purpose AI models, which accord-

until August 2027.¹⁴⁴ The timeline for enforcement is designed to give GPAI companies time to bring models into compliance and to develop the framework necessary to prove compliance.¹⁴⁵ Perhaps the timeline is justified by current model capabilities.¹⁴⁶ However, given the rapid advancement of AI over the last six years, any prediction about the intelligence and safety of models a year from now is difficult to trust. Some experts believe ASI will be developed before the COP is enforced.¹⁴⁷ Such a delay in enforcement could cause harm because, as noted in Part B.II, it may be impossible to align ASI after its development. Further, despite current models' capability shortcomings, we have already seen their ability to democratize dangerous CBRN knowledge.¹⁴⁸

2 Quasi Self-Regulation

Despite the strong language from the EU Commission, unlike the AI Act, the COP is not hard law. The COP offers guidance and is the preferred mechanism for demonstrating adherence to the AI Act, but it remains voluntary. GPAI companies are free to interpret the AI Act and show adherence on their own.¹⁴⁹ Without a detailed, mandatory regulatory framework, companies may choose to litigate their adherence to the AI Act. GPAI companies are flush with capital and will have little difficulty meeting the legal costs of resisting strict and timely regulation.¹⁵⁰

Further, even for signatories intending to follow the COP, there is plenty of self-regulation. For starters, the mandatory safety and security framework is developed independently by each company.¹⁵¹ Google's own framework discloses an intention to implement safety protocols only to the same extent as its competitors.¹⁵² "These recommended security levels....may be adjusted....if a model does not possess capabilities meaningfully different from other publicly available models that have weaker security applied (in which case the marginal benefit of higher security is limited)."¹⁵³ The message seems to be: we will play by the rules as long as it does not disadvantage us. The flexibility in the Safety and Security

ingly falls in with the Regulation's general application date of 2 August 2026); European Commission, *The Enforcement Framework of the AI Act*, available at <https://digital-strategy.ec.europa.eu/en/policies/enforcement-ai-act> ("The enforcement powers of the AI Office and the national competent authorities of the Member States apply from 2 August 2026").

144 Implementation Timeline, *supra* note 143.

145 *Overview of the Code of Practice*, EU Artificial Intelligence Act (30 July 2025), available at <https://perma.cc/53Y6-NV4S>.

146 Starr, *supra* note 33; Titus, *supra* note 5.

147 Roose, 'This A.I. Forecast Predicts Storms Ahead', *New York Times*, 3 April 2025, available at <https://perma.cc/J6HZ-BU54>.

148 Service, *supra* note 13.

149 AI Act, *supra* note 8, Art. 54(4).

150 Sigalos, 'OpenAI Wraps \$6.6 Billion Share Sale at \$500 Billion Valuation', *CNBC*, 2 October 2025, available at <https://perma.cc/XMV2-6Z24>.

151 GPAI Code of Practice, *supra* note 8, Commitment 1.

152 Google DeepMind, *Frontier Safety Framework, Version 3.0* (22 September 2025), available at <https://perma.cc/GK45-62K4>.

153 Frontier Safety Framework, *supra* note 153, at 9.

Framework leverages the in-house talent at each GPAI company. Such leverage is necessary as regulatory agencies cannot match the expenditure of these private companies. However, flexibility also leads to competition, as companies find ways to slightly advance their position by undermining safety.

3 *External Evaluation Gaps*

Supporters of the Code praise its “mandate” for independent evaluations of systemic risk and safety outlined in Appendix 3.5. However, the Code effectively nullifies this requirement by excusing companies from compliance when “[s]ignatories fail to appoint adequately qualified independent external evaluators despite using early search efforts”.¹⁵⁴ These “early search efforts” are defined as a public call open for just 20 business days.¹⁵⁵ Given the scarcity of properly credentialed external evaluators, allowing companies to forgo third-party review after so limited a search renders the supposed “mandate” largely meaningless. Further, the AI Act itself does not mandate third-party evaluations.¹⁵⁶ Because the AI Act itself does not mandate third-party evaluation, and AI companies are free to interpret the Act themselves, the enforcement of the Code’s mandate may be challenged in court.¹⁵⁷

4. *Transparency and Lifecycle Gaps*

The development of increasingly capable AI models will lead to major societal shifts.¹⁵⁸ The dangers include cementing institutional bias, massive job loss, and catastrophic loss of life.¹⁵⁹ Without transparency, the public cannot form an intelligent opinion on how AI should be deployed, advocate for desired policies, or ensure regulators are acting appropriately.

While the COP’s transparency requirements go further than comparable regulations, there are serious gaps. Commitment 10.2 states, “[i]f and insofar as necessary to assess and/or mitigate systemic risks, Signatories will publish (e.g. via their websites) a summarised version of their Framework and Model Report(s)”.¹⁶⁰ The COP does not explain what a summary must contain or when it is necessary to mitigate systemic risk. Extending redactions beyond sensitive commercial information could undermine the purpose of a public transparency commitment

154 GPAI Code of Practice, *supra* note 8, Appendix 3.5.

155 GPAI Code of Practice, *supra* note 8, Appendix 3.5.

156 *See* AI Act.

157 GPAI Code of Practice, *supra* note 8, Art. 54(4).

158 Fagan, *AI’s Impacts on Society: A Look into the Crystal Ball*, Mossavar-Rahmani Center for Business and Government, Harvard Kennedy School (2025), available at <https://perma.cc/7366-UFKW>.

159 Fagan, *supra* note 159.

160 GPAI Code of Practice, *supra* note 8, Commitment 10.2.

and rob society of its necessary input into decisions that will radically alter the future.

Furthermore, there is no requirement in the Code that serious incident reports be made publicly available. The Code mandates Model Reports to be submitted to the AI Office:

- Before Initial Market Placement¹⁶¹
- In Case of a Material Change to a Model Already on the Market¹⁶²
- Upon Request by the AI Office¹⁶³
- If a Serious Incident Mandates Reevaluation¹⁶⁴

These triggers limit the oversight of evolving in-house models. Recall that models under development have a greater risk of existential harm because engineers push and test model capabilities during development.¹⁶⁵ Because model reports are only mandatory prior to a model's public release, the only sources of knowledge of systemic risk in developing models are a serious incident report or a request from the AI Office.¹⁶⁶ Thus, despite the mandate to document, analyze, and justify systemic risks, the public will not influence development decisions, including when model development continues despite serious incidents in training. These loopholes not only keep indications of risk out of the public forum but also limit AI Office oversight until a serious incident has already occurred.

These internal gaps would be serious even if the COP applied universally. They become potentially dispositive when combined with the framework's limited reach. The most capable AI developers are not European, and the geopolitical forces driving them are not answerable to the AI Office.

II Application to Chinese and U.S. Developers

Frontier AI companies most capable of developing powerful, dangerous AI models are headquartered and operated out of the United States and China.¹⁶⁷ For the EU GPAI Code of Practice to avert AI-caused catastrophe, its safety requirements must reach beyond its borders to hold these developers accountable. The COP faces three interconnected challenges which are likely to limit its ability to regulate companies outside of Europe: (1) the potential divisibility of AI models, (2) Europe's diminished leverage over frontier AI companies, and (3) the geopolitical arms race incentivizing both the U.S. and China. The signatory position sharpens the first of these. The one major American developer to reject the Code outright, Meta, is also the one with the largest open-weight programme, the distribution channel to which the CBRN argument in Part B.I.1 is most exposed. Open model weights, once

161 GPAI Code of Practice, *supra* note 8, Measure 7.1.

162 GPAI Code of Practice, *supra* note 8, Measures 7.5-7.6.

163 GPAI Code of Practice, *supra* note 8, Commitment 10.2.

164 GPAI Code of Practice, *supra* note 8, Measure 10.1.

165 Anderljung et al., *supra* note 5.

166 GPAI Code of Practice, *supra* note 8, Commitment 10.2.

167 Emberson, *Chinese AI Models Have Lagged the US Frontier by 7 Months on Average Since 2023*, Epoch AI (2 January 2026), available at <https://perma.cc/33TF-N7B4>.

released, cannot be recalled by any regulator.¹⁶⁸ The opposite point should be conceded: the Safety and Security Chapter analysed here is the only chapter xAI did sign, so on this narrow question adherence is broader than the general picture of resistance suggests. That concession is bounded by the feature that runs through this Part: signing a voluntary code is not the same as being subject to an enforceable obligation.¹⁶⁹

1 *The Divisibility of AI Models*

Because international companies often choose to follow the strictest regulation to ensure compliance of their products worldwide, the AI Act and COP could govern models outside of the EU jurisdiction.¹⁷⁰ This de facto regulation by the EU is known as the “Brussels Effect”.¹⁷¹ This effect is bolstered by the size of the EU market. As the second-largest market in the world, frontier AI companies are unlikely to completely remove their models from the EU and thus forego revenue from a large market.¹⁷² Additionally, some argue that frontier AI companies are unlikely to develop separate models (that is, an EU-compliant model and a different model for other jurisdictions) due to the high cost of developing AI models.¹⁷³

In the past, EU regulations have set a global standard because firms have been unable to divert or separate their products from the EU market.¹⁷⁴ The EU’s General Data Protection Regulation (GDPR) became the global regulatory standard and was adopted by giant digital companies like Apple and Google, because of the “relative indivisibility of dataflows”.¹⁷⁵ Modern digital businesses often handle data from both EU and U.S. consumers on the same platforms.¹⁷⁶ Separating EU citizens’ data from U.S. citizens’ data would be virtually impossible.¹⁷⁷ Thus, companies decided to apply the EU regulation to data from citizens outside the EU jurisdiction.¹⁷⁸ The comparison should not be pressed too far. Where separation has been possible, firms have separated: Meta declined to release its multimodal Llama model in the European Union in July 2024, citing regulatory uncertainty,

168 Gade, Lermen, Rogers-Smith and Ladish, *BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B*, arXiv:2311.00117 (2023); Lermen and Rogers-Smith, *Badllama 3: Removing Safety Finetuning from Llama 3 in Minutes*, arXiv:2407.01376 (2024); AI Act, *supra* note 8, Art. 53(2) (open-source exemption inapplicable to models with systemic risk).

169 See *supra* note 61.

170 Siegmann and Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, Centre for the Governance of AI (August 2022), available at <https://perma.cc/4JGE-AYKK>, at 29.

171 Siegmann and Anderljung, *supra* note 171.

172 Siegmann and Anderljung, *supra* note 171, at 18-20.

173 Hoffmann, *AI Safety under the EU AI Code of Practice – A New Global Standard?*, Center for Security and Emerging Technology (30 July 2025), available at <https://perma.cc/K7V7-VJYC>.

174 Crum, ‘Brussels Effect or Experimentalism? The EU AI Act and Global Standard-Setting’, 14 *Internet Policy Review* (2025), issue 3, at 3.

175 Crum, *supra* note 175, at 4.

176 Voss, ‘Cross-Border Data Flows, the GDPR, and Data Governance’, 29 *Washington International Law Journal* (2020) 485.

177 Voss, *supra* note 177.

178 Voss, *supra* note 177.

Apple withheld Apple Intelligence from the European market on Digital Markets Act grounds, and Google released AI Overviews in the United States in May 2024 but had extended it to only eight Member States by April 2025, attributing the lag to European regulation.¹⁷⁹

Whether frontier AI models will be incapable of differentiation (one model for Europe and another for the world) is unclear.¹⁸⁰ Frontier AI companies may turn off certain model features in Europe to comply with the Code of Practice while allowing them to continue elsewhere.¹⁸¹ Additionally, developers may release older, less capable models in Europe while pressing ahead with cutting-edge abilities elsewhere.¹⁸² The potential divisibility of AI models is likely to limit the GPAI COP's impact outside the EU and thus have a limited effect in reducing the dangers of powerful AI systems.

2 Europe's Potential Lack of Leverage

Even if AI models are indivisible, Europe must have sufficient market leverage to enforce compliance. The EU relies on non-EU countries for over 80% of digital products, services, infrastructure, and intellectual property.¹⁸³ Europe's dependence on U.S. AI infrastructure is just as stark. Europe is estimated to account for only 7% of global compute.¹⁸⁴ The leading European AI company, Mistral, passed \$400 million in annualised revenue by January 2026 – a twentyfold increase that still leaves it more than an order of magnitude behind a single American developer.¹⁸⁵ OpenAI, by comparison, reported annualised revenue of \$40 billion by August 2026.¹⁸⁶ Europe is unlikely to be capable of challenging this lead as the talent pipeline and lack of investment exacerbate their problems. In 2025, 62% of EU AI post-docs said they planned to move to the U.S. or China, drawn by larger projects, higher salaries, and broader opportunities.¹⁸⁷ In 2025, 81% of private investment, or \$109 billion, went to U.S. AI firms versus only \$8 billion to European firms.¹⁸⁸

179 Gain, 'Meta Will Not Launch Multimodal Llama AI Model in EU', *Silicon Republic*, 18 July 2024, available at <https://www.siliconrepublic.com/business/meta-llama-3-multimodal-ai-model-llm-eu-europe-launch-delay>; Davies, 'Why OpenAI's Voice Mode, Meta's Llama and Apple's AI Won't Be Coming to Europe Yet', *Euronews*, 8 October 2024, available at <https://www.euronews.com/next/2024/10/08/why-openais-voice-mode-metas-llama-and-apples-ai-wont-be-coming-to-europe-yet>; Kroet, 'Google's AI Feature on Hold in Most EU Member States due to "Strict Rules"', *Euronews*, 1 April 2025, available at <https://www.euronews.com/next/2025/04/01/googles-ai-feature-on-hold-in-most-eu-member-states-due-to-strict-rules>; cf. Siegmann and Anderljung, *supra* note 171, at 39-42.

180 EJLR_24-25_188718

181 Siegmann and Anderljung, *supra* note 171, at 39-42.

182 Crum, *supra* note 175, at 11.

183 Swain, 'Europe Votes to Tackle Deep Dependence on US Tech in Sovereignty Drive', *Computerworld*, 23 January 2026, available at <https://perma.cc/8SEF-ZR9J>.

184 Kodama, *supra* note 59.

185 Epoch AI, *AI Companies* (updated 2026), available at https://epoch.ai/data/ai_companies_revenue_reports.csv.

186 Epoch AI, *supra* note 186.

187 Bairam, Lambert and Isabelle, 'Europe's AI Journey: From Dependency to Sovereignty', *European Economics*, 21 October 2025, available at <https://perma.cc/M9UA-ESXP>.

188 Matthews, 'State of AI 2025: Five Key Charts for Europeans', *Science Business*, 16 October 2025, available at <https://perma.cc/XC7B-T5VF>.

Given AI's paradigm-shifting ability and Europe's lack of home-grown AI infrastructure, the EU will be forced to choose between adopting U.S. or Chinese AI infrastructure.¹⁸⁹ The EU is unlikely to transition to a Chinese AI infrastructure due to security and human rights concerns.¹⁹⁰ Thus, the EU may need U.S. frontier AI companies more than such companies need the European market. This imbalance gives U.S. frontier AI companies significant leverage and the potential ability to skirt the EU's regulatory regime. Indeed, frontier AI companies appear willing to relax their own safety commitments where competition demands it: Anthropic's Responsible Scaling Policy version 3.0, effective February 2026, removed the commitment never to release catastrophically harmful AI.¹⁹¹ Over the same period Anthropic refused the Department of War's demand for authorisation to use its models for lethal autonomous warfare and mass domestic surveillance, was designated a supply-chain risk in consequence, and is litigating the designation.¹⁹² That is a firm choosing which constraints to keep under simultaneous competitive and political pressure: a sharp illustration of the limits of self-regulation.

3 The AI Arms Race

The development of powerful AI systems will have radical consequences across every societal domain. As with any powerful technology, AI will be applied for military advantage. In his thorough report on the development and implications of

189 Chhangani and Kumar, *What Drives the Divide in Transatlantic AI Strategy?*, Atlantic Council (29 September 2025), available at <https://perma.cc/GDF6-6FD7> ("At the software level, European firms overwhelmingly depend on US-developed foundational models, cloud platforms, and AI tools provided by companies such as Microsoft, Google, and OpenAI, reflecting the absence of a globally competitive European alternative.").

190 Liyanage, 'Tyranny of City Brain: How China Implements Artificial Intelligence to Upgrade Its Repressive Surveillance Regime', 4 *Journal of Illiberalism Studies* (2024) 73, available at <https://www.illiberalism.org/wp-content/uploads/2025/01/Chamila-Liyanage-Tyranny-of-City-Brain-1.pdf>.

191 Blair et al., *Department of War, Anthropic, and National Security*, AI Safety Newsletter No. 69, Center for AI Safety (13 March 2026), available at <https://perma.cc/U6TY-V4XU> ("Competitive pressures are creating a race to the bottom on frontier AI safety. Anthropic's justification for the changes are largely based on the fact that other AI companies are not going to stop development; the argument is that, if Anthropic alone were to stick to stricter safety commitments, it would simply fall behind other developers, while doing little to reduce overall risk. This causes a vicious cycle, as loosened safety commitments increase the speed of AI development, which in turn incentivizes further loosening."); Anthropic, *Responsible Scaling Policy*, Version 3.0 (effective 24 February 2026), available at <https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0>; Williams and Freund, *Anthropic's RSP v3.0: How It Works, What's Changed, and Some Reflections*, Centre for the Governance of AI (2026), available at <https://www.governance.ai/analysis/anthropics-rsp-v3-0-how-it-works-whats-changed-and-some-reflections>.

192 Congressional Research Service, *Pentagon-Anthropic Dispute over Autonomous Weapon Systems: Potential Issues for Congress*, IN12669 (2026); see *Anthropic PBC v. United States Department of War*, No. 3:26-cv-01996 (N.D. Cal. 26 March 2026) available at <https://law.justia.com/cases/federal/district-courts/california/candce/3:2026cv01996/465515/134/> (order granting preliminary injunction finding a high likelihood of success on Anthropic's First Amendment retaliation claim, and that the defendants had 'as of yet failed to prove that Anthropic's conduct qualifies as a supply-chain risk'); *Anthropic PBC v. United States Department of War*, No. 26-1049 (D.C. Cir. 2026) (per curiam order denying an emergency stay of the designation and expediting review).

ASI, *Situational Awareness – The Decade Ahead*, Leopold Aschenbrenner argues that superintelligence will bring

“[M]ilitary technological advances of a century compressed into less than a decade. We’ll see superhuman hacking that can cripple much of an adversary’s military force, roboarmies and autonomous drone swarms, but more importantly, completely new paradigms we can’t yet begin to imagine, and the inventions of new WMDs with thousandfold increases in destructive power (and new WMD defenses too, like impenetrable missile defense, that rapidly and repeatedly upend deterrence equilibria.)”¹⁹³

Aschenbrenner is not alone. In his report, *Promethean Rivalry – The World-Altering Stakes of Sino-American AI Competition*, Bill Drexel writes:

“[t]he transformative potential for AI in war is difficult to overstate, with an impact likely more significant than the advent of nuclear weapons ... AI systems offer faster, more precise, and more strategic capabilities across a far more expansive set of use cases ... AI will evolve novel varieties of warfighting ... that could create severe capability offsets difficult or impossible to match.”¹⁹⁴

The possessor of superintelligence will not only be able to develop new, powerful weapon systems but also build them at a rapid pace via automated robot factories.¹⁹⁵ This presumed power of ASI, and its influence on the global strategic balance, has led both the United States and China to declare their intent to “win” the AI race.¹⁹⁶

Due to the potential power of ASI, experts across disciplines, from AI engineering to international relations, have vigorously advocated for the United States’ preeminence in the AI race.¹⁹⁷ Such experts argue that only through achieving ASI first, with its decisive military and economic advantages, can the United States ensure the survival of liberal democracy.¹⁹⁸ These same experts argue that a decisive American lead is not merely a strategic goal but, paradoxically, a prerequisite for safe AI development – the theory being that only a country secure in its dominance can afford to slow down when serious dangers arise.¹⁹⁹

In this context, any EU regulation that impedes the development of powerful AI is likely unacceptable to the Trump Administration.²⁰⁰ Indeed, the U.S. government is vocal about its unhappiness with the AI Act and COP. In April 2025,

193 Aschenbrenner, *supra* note 5, at 129.

194 Drexel, *supra* note 42, at 10.

195 Aschenbrenner, *supra* note 5, at 129.

196 Executive Office of the President, *Winning the Race: America’s AI Action Plan* (July 2025), available at <https://perma.cc/9BE8-KEF8>, at 1; Webster et al., *China’s Plan to ‘Lead’ in AI: Purpose, Prospects, and Problems*, New America (1 August 2017).

197 Aschenbrenner, *supra* note 5, at 126; Drexel, *supra* note 42.

198 Aschenbrenner, *supra* note 5, at 134-136.

199 Aschenbrenner, *supra* note 5, at 127.

200 Perlo, ‘White House Tech Chief Slams EU AI Act, Champions Trump’s Approach as Davos Begins’, *NBC News*, 20 January 2026.

the Trump Administration sent a letter to the European Commission demanding an abandonment of the GPAI Code of Practice, viewing it as excessively burdensome.²⁰¹ At the World Economic Forum, the director of the White House Office of Science and Technology Policy, Michael Kratsios, called the AI Act an “absolute disaster”.²⁰² The U.S. refused to sign the international agreement on AI at the global summit in Paris, with Vice President Vance describing the document as “killing[ing] a transformative industry just as it’s taking off”.²⁰³ In a February 2025 memo, President Trump characterized EU digital rules and taxes as “overseas extortion” and “compromis[ing] American economic and national security interests”.²⁰⁴ This pressure campaign worked with the European Commission, “assessing changes intended to ease requirements for major tech companies”.²⁰⁵

That assessment is now complete. It produced the Digital Omnibus, on which political agreement was reached on 6 May 2026 and which Member State representatives confirmed on 13 May 2026; it deferred the high-risk obligations and simplified documentation for smaller providers. It did not touch the general-purpose AI chapter, the Article 50 transparency duties or the prohibited-practices regime. The Union deregulated at the margin under precisely the pressure described above, and left the provisions with which this Note is concerned exactly where they were.²⁰⁶ The lack of substantive changes to the GPAI COP indicates either a serious willingness to enforce it or that frontier AI companies do not view it as a substantial obstacle. Either way, the reach problem examined below is untouched.

Further, the EU’s AI regulation is unlikely to significantly burden Chinese developers. Chinese models, such as DeepSeek and Qwen, are not widely used in the EU market, with U.S. models commanding an 80% market share.²⁰⁷ The AI Act and the COP are correspondingly easier for the ruling Chinese Communist Party (CCP) and private Chinese AI companies to ignore. Nor is the CCP likely to abandon its use of AI for facial recognition, mass surveillance and social control in order to

201 Volpicelli, Stolton and Nardelli, ‘Trump Administration Pressures Europe to Ditch AI Rulebook’, *Bloomberg*, 26 April 2025, available at <https://www.bloomberg.com/news/articles/2025-04-25/trump-administration-pressures-europe-to-reject-ai-rulebook>.

202 Perlo, *supra* note 201.

203 Kleinman and McMahon, ‘UK and US Refuse to Sign International AI Declaration’, *BBC News*, 11 February 2025, available at <https://perma.cc/NH9P-GNLE>.

204 The White House, *Memorandum on Defending American Companies and Innovators from Overseas Extortion and Unfair Fines and Penalties* (21 February 2025), available at <https://perma.cc/LD25-T54X>, para. 1.

205 *EU Weighs Delaying Key AI Act Rules after Pressure from Trump Administration, Tech Companies*, Anadolu Agency, 8 November 2025, available at <https://perma.cc/3UBW-74DP>.

206 Regulation amending Regulation (EU) 2024/1689 (Digital Omnibus on artificial intelligence), provisional political agreement of 6 May 2026, confirmed by Member State representatives on 13 May 2026; see Gibson Dunn, *EU AI Act Omnibus Agreement – Postponed High-Risk Deadlines and Other Key Changes* (2026), available at <https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/>; see European Commission, *supra* note 144 (on GPAI unchanged timetable).

207 Balland et al., *Generative AI and Foundation Models in the EU: Uptake, Opportunities, Challenges, and a Way Forward*, Centre for European Policy Studies (30 January 2026), available at <https://perma.cc/9DRR-NF8Q>.

comply with EU regulation.²⁰⁸ Lastly, like the U.S., China sees AI development as a strategic necessity.²⁰⁹ Given these factors, Chinese AI developers have little incentive to adhere to the COP.

The COP's inability to control Chinese AI development effectively gives the US a further incentive to oppose these EU rules.²¹⁰ Because the Trump Administration views the COP as a strain on innovation, and given the perceived strategic necessity of AI leadership, full adherence to the AI Act and COP would amount to unilateral disarmament. Given these incentives, it is unlikely that the US will adhere to the safety and security requirements set out in Chapter Three of the GPAI Code of Practice.

To avert the potential catastrophic disasters described in Part B.I, the EU GPAI Code of Practice must have a global regulatory effect. Given frontier AI companies' potential ability to adhere to the COP while rapidly pursuing ASI (via model differentiation), Europe's lack of AI infrastructure options, and the AI arms race, it appears unlikely the COP will have a significant Brussels Effect. Should the COP regulate only AI development within its borders, numerous companies will continue to race ahead in developing superintelligence, with little attention paid to safety. Once created, an unsafe ASI will not be confined to political borders.²¹¹ Without significant adherence by U.S. and Chinese AI companies outside EU jurisdiction, the COP cannot avert an AI-caused catastrophe.

E A Solution – International Safety Treaty

The analysis in the preceding sections reveals a fundamental mismatch between the scope of the problem and the reach of the solution. Despite the COP's strength, the companies most capable of producing an AI catastrophe are headquartered outside of Europe, governed by administrations that have publicly declared EU AI regulation an obstacle to national strategy, and operating in a geopolitical environment that treats safety constraints as a competitive handicap.²¹² The best answer, one on which legal and policy scholars across disciplines have increasingly converged, is an international treaty binding the major AI-developing powers to common safety standards.²¹³

As discussed in Part D.II.1, frontier AI companies may respond to EU regulation by simply turning off features for European users while continuing unrestricted development in their home jurisdictions. Unlike the GDPR, whose Brussels Effect

208 Liyanage, *supra* note 191.

209 Webster et al., *supra* note 197.

210 Baland et al., *supra* note 208.

211 Yudkowsky and Soares, *supra* note 5.

212 America's AI Action Plan, *supra* note 197.

213 Wasil et al., *International Agreements on AI Safety: Review and Recommendations for a Conditional AI Safety Treaty*, arXiv:2408.16074 (March 2025), available at <https://arxiv.org/html/2408.16074v2>; Dias, *AI Agents and Global Governance*, Partnership on AI (2025), available at <https://partnershiponai.org/wp-content/uploads/2025/09/agents-policy-analysis.pdf>; Maas and Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, Institute for Law and AI, AI Foundations Report No. 1 (2023), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4579773.

derived from the near-impossibility of separating European data from global data flows, AI models are architecturally separable.²¹⁴ A company can serve a restricted model to European markets and a full-capability model everywhere else. A treaty eliminates the problem entirely. If the United States and China are both signatories, there is no regulation-free jurisdiction to escape to. The race-to-the-bottom dynamic that currently makes safety constraints feel self-defeating is replaced by a level playing field: every major developer faces the same requirements, and no one gains a competitive advantage by cutting safety corners.²¹⁵

The leverage problem compounds the divisibility concern. Even if a company cannot simply turn off EU-restricted features without consequence, Europe currently lacks the infrastructure power to make that consequence credible. As documented in Part D.II.2, the EU accounts for only seven percent of global AI compute, depends on U.S. frontier models for over eighty percent of its digital AI services, and has seen the Trump Administration openly threaten retaliatory consequences for AI regulation.²¹⁶ A bilateral or multilateral treaty reframes the entire dynamic. The EU is no longer a lone regulator demanding compliance from companies that can simply threaten to leave. It becomes a co-equal party, alongside the United States and China, to an agreement that neither great power can escape without consequence. Europe's weakness in the unilateral regulatory context becomes irrelevant when the major powers have jointly accepted the same obligations.

The arms race problem is the deepest, but it too has a treaty solution. The United States has characterized winning the AI race as a strategic national security necessity akin to the development of nuclear weapons.²¹⁷ China has declared a near-identical goal.²¹⁸ In this context, any safety regulation that applies only to one side is strategically disadvantageous. However, China and the United States have shared interests in safe development of AI.²¹⁹

An agreement on the safe innovation of AI will allow the trade relationship between the US and China to continue.²²⁰ As AI diffuses into more products, suspicion of imports will continue to rise.²²¹ This suspicion will not be held solely by intelligence agencies but by average consumers. The suspicion of foreign AI-enabled devices could radically reduce trade.²²² The battles over the ownership

214 See supra notes 175–183 and accompanying text; Crum, supra note 175.

215 Chase and Marcellino, *Potential for U.S.-China Cooperation on Reducing AI Risks*, RAND (August 2025), available at <https://www.rand.org/pubs/perspectives/PEA4189-1.html>; Chan and Hass, *Can the US and China Cooperate on AI?*, Brookings (May 2026), available at <https://www.brookings.edu/articles/can-the-us-and-china-cooperate-on-ai/>; McGuire, *How Trump Should Approach AI Talks with China*, Council on Foreign Relations (May 2026), available at <https://www.cfr.org/articles/how-trump-should-approach-ai-talks-with-china-targeted-dialogue-maximum-pressure>.

216 See supra notes 203–206.

217 America's AI Action Plan, supra note 197.

218 Webster et al., supra note 197.

219 See supra note 216.

220 Friedman, 'The One Danger That Should Unite the U.S. and China', *New York Times*, 2 September 2025, available at <https://www.nytimes.com/2025/09/02/opinion/ai-us-china.html>.

221 Friedman, supra note 221.

222 Friedman, supra note 221.

of TikTok are an early example of how difficult trade will become without trust in the product.²²³ An agreement to embed ethics and safety into every AI model allows a thriving trade relationship to continue.

Both governments also share a fundamental interest in keeping powerful AI out of the hands of criminals and non-state actors.²²⁴ Powerful AI without proper safety limitations will empower the worst criminals and actors in every society.²²⁵ Ensuring law enforcement's ability to keep these powerful technologies out of the hands of criminals is surely a goal of both the United States and China. However, without agreement between powers, bad actors will simply seek out and use the least secure models.²²⁶ Their actions are unlikely to be confined by political borders.

Perhaps most fundamentally, both governments share an interest in preserving their own authority. Some experts believe that the development of superintelligence could empower the corporation above the nation-state.²²⁷ Both governments will be driven to empower corporate actors in the pursuit of winning the arms race. Once in the possession of superintelligence, such corporations "could launch cyberattacks, wield political influence, or fashion new and exotic weapons to seize power".²²⁸ Here again, a treaty on the proper development and safety of AI is in the best interests of both the United States and China.

History offers a clear precedent for breaking the arms-race dynamic. In the 1970s, the United States and Soviet Union were engaged in active global competition, including proxy conflicts on multiple continents.²²⁹ Nevertheless, they reached the Strategic Arms Limitation Agreements and the Anti-Ballistic Missile Treaty, not because they trusted each other, but because both recognized that unconstrained competition in the most dangerous domain was not in either party's interest.²³⁰ The same logic applies here. A misaligned superintelligence will not observe political borders.²³¹ While American and Chinese policymakers may compete fiercely for other advantages, they share a common interest in not being killed by their own creation.

However, precedent and an appeal to shared interests alone may not be enough. Part D established that the United States treats safety constraints as unilateral disarmament and that China has little reason to comply with rules it did not write. The position has hardened since: Executive Order 14365 of 11 December 2025 directs the Attorney General to establish a task force to litigate against state AI

223 Friedman, *supra* note 221.

224 Chan and Hass, *supra* note 216.

225 Chan and Hass, *supra* note 216.

226 Chan and Hass, *supra* note 216 ("Common standards would help prevent 'safety arbitrage,' where malicious actors simply seek out the least restrictive model available in either country.").

227 Segal, *The AI Balance of Power*, Council on Foreign Relations (20 May 2026), available at <https://www.cfr.org/articles/the-ai-balance-of-power>; Davidson, *How AI Could Help Overthrow Governments*, Future of Life Institute Podcast (17 July 2025), available at <https://futureoflife.org/podcast/how-ai-could-help-overthrow-governments-with-tom-davidson/>.

228 Segal, *supra* note 228.

229 Thomas, 'Synergy in Paradox: Nixon's Policies toward China and the Soviet Union', 5 *Chicago Journal of History* (2015) 29, available at <https://cjh.uchicago.edu/issues/fall15/5.3.pdf>.

230 Thomas, *supra* note 230.

231 See generally *supra* note 214, 216.

laws, conditions federal assistance on state legislative alignment, and requires the preparation of a federal preemption bill.²³² An administration hostile even to domestic regulation is a harder audience for an international instrument, and a more urgent reason to seek one.

The same incentives that doom the COP would seem to doom any treaty built upon it: why would two parties in a declared race sign and adhere to a binding safety agreement? The most recent evidence is not encouraging. The New Delhi Declaration adopted at the India AI Impact Summit in February 2026 was endorsed by ninety-one countries and international organisations, including both the United States and China – the same United States that had refused the Paris declaration a year earlier.²³³ It is expressly non-binding and contains no safety commitment capable of constraining a signatory.²³⁴ When the two rivals do sign the same document, it is because the document asks nothing of them.

The candid answer is not that common interest will manufacture mutual trust: it will not, and arms-control scholarship warns that the nuclear analogy can mislead as readily as it can instruct. There are fundamental differences: AI is a general-purpose, easily replicated technology rather than a discrete weapon kept under government lock and key.²³⁵ What the Cold War record shows is narrower and more useful: rivals who fully expected continued competition still sustained agreements when adherence could be observed rather than trusted. The 1963 hotline and the 1972 Incidents at Sea Agreement, concluded alongside SALT I, held not because Washington and Moscow trusted each other but because each could watch and verify treaty obligations.²³⁶

Verification, not trust, is what held those agreements together, and frontier AI's reliance on physical, trackable infrastructure is what makes verification possible here. An international AI safety body, staffed by technical experts from signatory states and empowered to conduct independent model evaluations, could do for AI governance what the IAEA does for nuclear nonproliferation: create a standing regime of mandatory reporting, third-party evaluation, and binding compliance decisions that no single jurisdiction can provide on its own.²³⁷ Something of this kind has begun to exist. United Nations General Assembly resolution A/RES/79/325 of 26

232 Executive Office of the President, Exec. Order No. 14365, *Ensuring a National Policy Framework for Artificial Intelligence*, 11 December 2025, available at <https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/>.

233 Ministry of External Affairs, Government of India, *AI Impact Summit 2026 Concludes with Adoption of New Delhi Declaration* (February 2026); Press Information Bureau, *More Countries Join the New Delhi Declaration on AI Impact* (24 February 2026) (recording ninety-one endorsing countries and international organisations); See Kleinman and McMahon, *supra* note 204.

234 Priyadarshini, 'Global Leaders Endorse Delhi Declaration on Safe and Responsible AI', *Nature*, 26 February 2026, available at <https://www.nature.com/articles/d44151-026-00036-6> (describing the declaration as non-binding).

235 McNamara, *Nuclear Arms Control Policies and Safety in Artificial Intelligence: Transferable Lessons or False Equivalence?*, Finnish Institute of International Affairs Briefing Paper, available at <https://fia.fi/en/publication/nuclear-arms-control-policies-and-safety-in-artificial-intelligence>.

236 Horowitz and Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, Center for a New American Security (January 2021), available at <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>, at 11.

237 Ansari, *Hardware-Level Governance of AI Compute*, arXiv:2604.04712 (April 2026), available at <https://arxiv.org/abs/2604.04712>.

August 2025 established the Independent International Scientific Panel on AI together with the Global Dialogue on AI Governance, which first convened in the margins of the eightieth session in September 2025 and meets annually. But the Panel assesses and the Dialogue convenes; neither inspects, and neither can find a provider non-compliant.²³⁸ The machinery the world has built so far is the machinery of dialogue, which is precisely the distinction drawn here.

Scholars have proposed that compute thresholds offer a practical basis for treaty oversight, analogous to the quantities of fissile material that trigger IAEA inspection.²³⁹ Frontier AI training runs at the scale relevant to existential risk require massive, physically detectable infrastructure: chips manufactured by fewer than five companies globally, power consumption measurable from the outside, and supply chains that pass through allied nations' jurisdictions.²⁴⁰ The chokepoints are specific. Extreme ultraviolet lithography systems, without which leading-edge logic cannot be fabricated at all, are supplied by a single firm in a single jurisdiction, and frontier accelerators are fabricated at a handful of leading-edge foundries. This is not a hypothetical control point: the United States and its partners already operate export controls at exactly this node, so a treaty verification regime would extend an enforcement mechanism that exists and functions rather than invent one.²⁴¹ These verification mechanisms are meant to function under rivalry rather than consensus.²⁴² Monitoring is difficult; it is not impossible.

Acknowledging the potential for verification does not settle the matter. A determined state could still cheat, and verification at the frontier will always trail the capabilities it polices. But the realistic choice is not between a perfect treaty and an imperfect one. It is between an enforceable floor that both powers can monitor and the arrangement Part D describes, in which the safety of humanity rests on the voluntary forbearance of companies sprinting toward superintelligence. A regime that renders cheating detectable, if not impossible, alters that calculus in a way the COP, anchored in self-assessment and a Brussels Effect, cannot.

If the case for an enforceable treaty holds, the question becomes what to build it from. Here, the negotiating parties need not start from a blank page. The GPAI

238 UN General Assembly Res. A/RES/79/325 (26 August 2025), *Terms of Reference for the Independent International Scientific Panel on Artificial Intelligence and Modalities for the Launch of the Global Dialogue on Artificial Intelligence Governance*, available at <https://docs.un.org/en/A/RES/79/325>.

239 Wasil et al., *supra* note 214; Harack et al., *Verification for International AI Governance*, Oxford Martin AI Governance Initiative, University of Oxford (2025), available at https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf.

240 Wasil et al., *supra* note 214.

241 Lin, *Beyond TSMC and EUV: Five Key Points on Chip Chokepoints*, Emerging Technology Observatory, available at <https://eto.tech/blog/five-key-facts-chokepoints-chip-supply-chain/> (reporting that extreme ultraviolet scanners are supplied entirely by a single Netherlands-based firm and that electronic design automation software is 96 per cent United States-supplied); Low, *Semiconductor Manufacturing Equipment Export Controls*, Institute for AI Policy and Strategy, available at <https://www.iaps.ai/research/semiconductor-manufacturing-equipment-export-controls>.

242 Siddik, *Toward ASI Stability: A Treaty Framework for US-China Cooperation on Artificial Superintelligence*, Convergence Analysis (2025), available at <https://www.convergenceanalysis.org/fellowships/international-security/toward-asi-stability-a-treaty-framework-for-us-china-cooperation-on-artificial-superintelligence>.

Code of Practice offers a ready framework for the core technical architecture of any such treaty. Its definition of dangerous model capabilities and propensities provides exactly the kind of precise, technically grounded scope and definition that any effective agreement must contain.²⁴³ Its safety and security framework, model reporting requirements, and lifecycle monitoring obligations demonstrate that rigorous, detailed AI safety regulation is achievable in practice.²⁴⁴ The COP is not the solution. It is the blueprint.

F Conclusion

The Safety and Security Chapter of the EU GPAI Code of Practice represents the most serious attempt yet to govern the development of powerful AI. By requiring providers to identify, analyze, and justify the systemic risks embedded in their models, the Code takes a genuine step toward meaningful oversight. Its enumeration of dangerous capabilities and propensities, its lifecycle monitoring requirements, and its whistleblower protections collectively establish a framework that is more rigorous and more specific than anything previously attempted.

But rigor on paper is not safety in practice. The COP's voluntary architecture, its loopholes for independent evaluation, and its vague transparency commitments leave the most consequential decisions in the hands of the very companies being regulated. More fundamentally, the Code will struggle to govern AI development outside European borders and the companies most capable of producing the dangerous models described in Part B are not European. As long as the United States treats safety regulation as a strategic liability and China treats it as irrelevant, the COP cannot close the gap between the world it envisions and the one it governs. The safety of humanity will not be determined by whether Anthropic files a thorough Model Report with the AI Office. It will be determined by whether the engineers racing toward superintelligence, in San Francisco and Beijing, are forced to stop when they find something dangerous. The EU GPAI Code of Practice, for all its genuine strengths, provides no such guarantee. Civilization races forward. Without a global AI governance treaty, the strategy, it seems, remains hope.

²⁴³ GPAI Code of Practice, *supra* note 8, Appendices 1.3-1.4.

²⁴⁴ *See supra* Part C.II-IV.