The Case for an AI Safety Safe Harbor

By Will Rinehart[1]

Last summer, OpenAI and Anthropic conducted the first-of-its-kind joint evaluation exercise where each company tested the other's publicly released AI models. The results, published in parallel in August 2025 by OpenAI and Anthropic, mark a significant moment in AI governance where competitors collaborated to test whether frontier AI systems might attempt harmful or deceptive actions.[2,3] The OpenAI and Anthropic project was limited but it's a sign that the next phase of AI governance is going to require that competitors coordinate on safety and security.

However, both companies are now wading into murky regulatory territory. OpenAI and Anthropic are competitors, and traditionally, the Federal Trade Commission and the Department of Justice have not looked favorably on information-sharing practices between rivals. Meanwhile, the withdrawal of the 2000 Antitrust Guidelines for Collaboration Among Competitors in December 2024 has left the legal landscape for such cooperation uncertain at exactly the moment when AI safety may demand it most.

Thus, as part of this inquiry, the Department of Justice (DOJ) and the Federal Trade Commission (FTC) should issue new collaboration guidance for AI safety, making clear that the agencies will not challenge properly designed AI safety collaborations. The final section of this filing includes a draft proposal. It is preceded by sections that detail the history of the guidelines and context surrounding AI safety.

A Brief History of Collaborations Among Competitors

In the waning weeks of the Biden Administration, the Federal Trade Commission and the Department of Justice took the unusual step of withdrawing the Antitrust Guidelines for Collaboration Among Competitors that had been in place since 2000.[4] In the two-page joint statement that was released on December 11, 2024, the agencies stated that these guidelines no longer provided reliable guidance due to evolving case law and changes in enforcement priorities.[5] Most notably, the joint statement singled out "rapidly changing technologies such as artificial intelligence [and] algorithmic pricing models" as reasons driving the withdrawal.

---

[1] The American Enterprise Institute (AEI) is a nonpartisan, nonprofit, 501(c)(3) educational organization and does not take institutional positions on any issues. The views expressed in this regulatory comment are those of the authors. This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

[2] *Findings from a pilot anthropic–openai alignment evaluation exercise: Openai Safety tests*. OpenAI. (2025, August 25). https://openai.com/index/openai-anthropic-safety-evaluation/

[3] Bowman, S. R., Srivastava, M., Kutasov, J., Wang, R., Bricken, T., Wright, B., Perez, E., & Carlini, N. (2025, August 27). *Findings from a pilot anthropic-openai alignment evaluation exercise*. Anthropic. https://alignment.anthropic.com/2025/openai-findings/

[4] *Antitrust guidelines for collaborations among competitors*. Federal Trade Commission. (2000, April). https://www.ftc.gov/sites/default/files/documents/public_events/joint-venture-hearings-antitrust-guidelines-collaboration-among-competitors/ftcdojguidelines-2.pdf

[5] *Justice Department and Federal Trade Commission withdraw guidelines for collaboration among competitors*. Federal Trade Commission. (2024, December 11). https://www.ftc.gov/system/files/ftc_gov/pdf/v250000collaborationguidelineswithdrawalstatement.pdf

Both Republican commissioners roundly criticized the move, arguing it was ill-timed in light of the impending change in administration. As Commissioner Holyoak noted in her dissent, the Democratic commissioners withdrew the Collaboration Guidelines without outlining any plans for their replacement which "leaves businesses grasping in the dark."[6] Commissioner Andrew Ferguson, who now serves as FTC Chair, in part agreed with the revocation of the guidelines, saying that the "Commission should from time-to-time revisit its nonbinding guidance to ensure that it properly informs the public of the Commission's enforcement position, promoting transparency and predictability."[7] But he too chided the late hour revision, arguing that "Now is the time to facilitate an orderly transition, not to withdraw existing guidance or to push through revised or new guidance."

While the decision to withdraw the 2000 Collaboration Guidelines affects all industries, it is particularly relevant for companies developing the most advanced AI systems. In January 2024, the FTC issued orders to Alphabet, Amazon.com, Anthropic, Microsoft and OpenAI to provide information on their recent investments and partnerships with cloud service providers.[8] The Friday before President Trump took office, the FTC released the 6(b) study, which was "aimed at helping the agency, the public, and policymakers deepen their understanding of the corporate partnerships formed between the generative AI developers and [cloud service providers] included in" the inquiry.[9] But in a largely unnoticed line on page 18, which discussed Microsoft's short-lived observer seat on OpenAI's board, a Microsoft employee said that they plan to continue holding "regular stakeholder meetings to share progress on our mission and ensure stronger collaboration across safety and security." In other words, even when AI companies faced elevated antitrust scrutiny, collaborations aimed at AI safety and security were still worth the risk.

Importantly, the Joint Withdrawal Statement stipulated in a footnote that the action did not affect the 2014 Antitrust Policy Statement on Sharing of Cybersecurity Information and stood by the assessment that "'properly designed sharing of cyber threat information should not raise antitrust concerns,' given the valuable purpose of sharing, and the very technical nature of, such information."[10] Still, the 2014 Statement focuses on traditional cybersecurity threats, but it doesn't address many of the critical AI safety concerns we face today like model weights security, adversarial attacks on AI systems, and alignment failures.

Congress has a history of providing certain antitrust protections for collaborative research and development in emerging technologies, as antitrust was viewed as a hindrance to American

---

[6] Holyoak, M. (2024, December 11). *Dissenting Statement of Commissioner Melissa Holyoak*. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/holyoak-collaboration-guidelines-withdrawal-statement.pdf

[7] Ferguson, A. (2024, December 11). *Dissenting statement of commissioner Andrew N. Ferguson*. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/collaborations-guidance-withdrawal-ferguson_-statement.pdf

[8] *Partnerships between cloud service providers and AI developers: FTC staff report on AI partnerships & investments 6(b) study*. Federal Trade Commission. (2025, January). https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf

[9] *Department of Justice and Federal Trade Commission: Antitrust policy*. Department of Justice. (2014, April 10). https://www.ftc.gov/system/files/documents/public_statements/297681/140410ftcdojcyberthreatstmt.pdf

[10] *Justice Department and Federal Trade Commission withdraw guidelines for collaboration among competitors*. Federal Trade Commission. (2024, December 11). https://www.ftc.gov/system/files/ftc_gov/pdf/v250000collaborationguidelineswithdrawalstatement.pdf

competitiveness in global markets. In 1993, Congress passed the National Cooperative Research and Production Act (NCRPA), an extension of the previous National Cooperative Research Act of 1984, to minimize the legal exposure of joint ventures under antitrust law. Then, in 2004, Congress passed the Standards Development Organization Act (SDOAA) to extend antitrust protections to standards development organizations.

The point of these statutes was not to exempt competitors from antitrust law altogether, but to reduce the legal uncertainty surrounding collaborative research and standards-setting. Congress understood that some forms of coordination are pro-consumer efforts to solve technical problems that no single firm can solve as efficiently on its own. An AI safety safe harbor would follow in that tradition.

Safety

The case for an AI safety safe harbor is built on practical reality. It is an increasingly technical field concerned with identifying failures in advanced AI systems, including jailbreaks, prompt injection, model-weight theft, training-data contamination, hallucination, sycophancy, deceptive behavior, cyber misuse, CBRN assistance, autonomous agent failures, and efforts by models to undermine oversight. These concerns are not the ordinary product-quality issues imagined by antitrust law. They are shared safety and security problems that cut across firms, models, and deployment environments.

The joint evaluation exercise between Anthropic and OpenAI underscores this point. The exercise focused on understanding model propensities, the kinds of concerning behaviors models might attempt in difficult scenarios, rather than estimating real-world likelihoods of harm. Both companies facilitated the evaluations by relaxing some external safeguards that would otherwise interfere with testing, a common practice for dangerous-capability assessments.

The findings were mixed but illuminating. No model from either developer was “egregiously misaligned,” but all exhibited concerning behaviors in stress-test environments, including cooperation with simulated misuse requests, sycophantic validation of delusional beliefs, and occasional attempts at whistleblowing or even blackmail when placed in extreme fictional scenarios. OpenAI's o3 reasoning model generally performed well on alignment dimensions, while Anthropic's Claude models showed strength in instruction-following but higher rates of refusing to answer uncertain questions.

Each lab has different internal evaluation tools, different safety cultures, different threat models, and different assumptions about where systems are likely to fail. A test designed by one lab can reveal weaknesses that another lab’s process might miss. Conversely, a model that performs well against one developer’s benchmarks may fail when evaluated under another developer’s scaffolding. In other words, safety evaluation is not merely a firm-specific compliance function but is fundamentally a shared measurement problem.

Still, the exercise was backward-looking and limited to publicly released models. As Nicholas Felstead convincingly argued in *Lawfare*, it could be because of the legal concerns: "this

collaboration focused on publicly released models — potentially due to the fear that collaborating on unreleased models would raise regulatory scrutiny."[11]

A one-off evaluation of public models is useful, but it is not enough for this fast-moving environment. To make it useful in practice, collaboration will need to happen earlier and under greater confidentiality through pre-deployment evaluation of unreleased models, coordinated vulnerability disclosure, exchange of information about security, and development of shared evaluation infrastructure, among others.

But these are precisely the agreements that can give antitrust enforcers anxiety because they require competitors to communicate about sensitive technical systems before public release. Indeed, Anthropic has filed before on this issue before, noting that, "Clarity on antitrust regulation would help determine whether and how AI labs can coordinate on safety standards. Sensible coordination around consumer-friendly standards seems possible, but regulators' guidance on the issue would be welcome."[12]

Furthermore, the work of the Frontier Model Forum is proof of this practice. In February 2026, FMF reported that its information-sharing work had enabled firms to exchange information about frontier-model vulnerabilities, exploitable flaws, adversarial inputs, data poisoning, threat actors, attack vectors, CBRN risks, offensive cyber capabilities, and model autonomy.[13] FMF also emphasized that its legal and technical infrastructure was designed to protect intellectual property and ensure antitrust compliance. The DOJ and the FTC should take steps to encourage these activities.

Draft Language for an AI Safety Safe Harbor

Properly understood, AI safety collaboration is a kind of precompetitive technical coordination. It helps firms discover and mitigate shared problems while also preserving competition over products, prices, customers, and capabilities. While the best pathway to ensuring collaboration would be a bill from Congress, the DOJ and the FTC could help provide clarity by issuing a joint policy statement setting out an AI safety safe harbor. A draft of that statement begins on the next page. It sets out terms for structured, technical, safety-focused collaboration while excluding prices, customers, output, wages, commercialization plans, and other competitively sensitive business information.

---

[11] Felstead, N. (2026, March 5). *How antitrust can promote AI safety collaborations*. Lawfare. https://www.lawfaremedia.org/article/how-antitrust-can-promote-ai-safety-collaborations

[12] *Charting a path to ai accountability*. Anthropic. (2023, June 13). https://www.anthropic.com/news/charting-a-path-to-ai-accountability

[13] *Progress update: FMF Information Sharing of Frontier AI threats and vulnerabilities*. Frontier Model Forum. (2026, February 16). https://www.frontiermodelforum.org/updates/progress-update-fmf-information-sharing-of-frontier-ai-threats-and-vulnerabilities/

An AI Safety Safe Harbor

The Department of Justice and the Federal Trade Commission, hereafter the Agencies, will not challenge a collaboration among actual or potential competitors under Section 5 of the FTC Act or Section 1 of the Sherman Act when all of the following conditions are satisfied. Covered collaborations will be evaluated under the rule of reason, not *per se* illegality.

Covered Purpose. The collaboration is reasonably necessary to identify, evaluate, disclose, or mitigate material safety or security risks arising from the development, release, or deployment of general-purpose or sector-specific AI systems. A risk is material if it creates a reasonable probability of significant harm to persons, critical infrastructure, or the integrity of AI systems themselves, including risks arising from misalignment, adversarial manipulation, model weight exfiltration, training data contamination, or loss of human oversight.

Covered Systems. The safe harbor applies to collaborations of foundation models, general-purpose AI systems, fine-tuned derivatives, and open-weight model releases. Coverage extends to both released and unreleased models where the collaboration is directed at safety evaluation rather than coordinating commercial release strategy.

Covered Activities. Covered activities include pre-deployment and post-deployment safety evaluations; evaluations of unreleased models conducted under confidentiality protocols; red-teaming and adversarial stress-testing; coordinated vulnerability disclosure; incident and hazard reporting; sharing technical indicators of model misuse, weight compromise, adversarial distillation attempts, training data contamination, or security incidents; development of safety benchmarks or interoperability protocols; joint research on technical mitigations, safeguards, watermarking, provenance systems, or containment tools; dataset protocols necessary to address safety risks; and security practices dealing with model weights and model provenance standards.

Participant Eligibility. Participants may include AI developers, cloud service providers, and research institutions with a demonstrated technical role in the covered safety activity. Foreign-domiciled entities may participate subject to applicable national security review. No participant may use membership in a covered collaboration to exclude a competitor from participation where that exclusion is not reasonably necessary to the safety objective. The collaboration must document the basis for participant eligibility restriction.

Necessary Limitation. The collaboration must be limited in scope, duration, and participants to what is reasonably necessary to achieve the safety objective. Any restraint on rivalry must be ancillary to a legitimate safety purpose rather than a naked agreement to limit competition. Participants may not use a covered collaboration as a vehicle to coordinate on matters outside the defined safety scope or those ancillary to it.

Excluded Subjects. The safe harbor does not apply to agreements over prices, output restrictions, customers, territories, wages or compensation, product commercialization strategy, nonpublic business plans, or other competitively sensitive information not reasonably necessary to the covered safety activity. Product commercialization strategy does not apply to joint safety outputs, such as shared benchmarks, evaluation tools, or incident databases, when those outputs are made available on non-discriminatory terms.

Intellectual Property. Participants must establish written agreements governing ownership, licensing, and publication rights for outputs produced by the collaboration before substantive work begins. An IP agreement that reserves commercialization rights to a single participant, or that restricts access to jointly developed safety tools in ways that competitively disadvantage other participants, may indicate that the collaboration extends beyond covered purposes.

Governance Safeguards. A covered collaboration must maintain written protocols specifying the collaboration's defined safety objective, covered activities, and excluded subjects; recordkeeping sufficient for antitrust review; and an antitrust compliance program.

Notice and Transparency. A collaboration seeking safe harbor protection shall file notice with both Agencies within 30 days of commencement of substantive collaboration. The notice shall include the identity of all participants; a description of the defined safety objective; a list of covered activities; a summary of the governance safeguards; and identification of any government agency or accredited evaluator involved. Material changes in participants or covered activities require supplemental notice within 30 days.

Reservation of Authority. The Agencies retain authority to challenge conduct used as a pretext for price fixing, market allocation, exclusionary standard manipulation, or other anticompetitive behavior. This statement does not alter the rights of private litigants or courts under otherwise applicable law and does not constitute agency approval of any specific collaboration.