

# AI Security Tax Incentive

The AI industry faces a critical misalignment of incentives. Commercial pressures often lead companies to prioritize rapid development and deployment over safety and security research, creating an *AI investment differential*.

Recent events at OpenAI illustrate this challenge. In 2023, OpenAI took a promising step, [announcing](#) a commitment to allocate 20% of its compute budget towards a new “superalignment” team focused on AI safety research. Less than a year later, this superalignment team was [disbanded](#) and its co-leads resigned.<sup>1</sup> Separately, in 2024 OpenAI’s [Preparedness Team reportedly](#) rushed safety evaluations for their flagship GPT-4o model’s launch to maintain a market advantage over competitors. A post-launch internal analysis found the model actually exceeded OpenAI’s own standards for harmful persuasion.

**To mitigate the incentive imbalance, we propose implementing a tax incentive structure that rewards sustained investment in AI safety and security.** This structure would help realign commercial interests with the public interest in maintaining safe and reliable AI systems.

## Core Components of Proposed AI Security Tax Incentive

### 1. *Tax Credit Structure*

- 25 percent tax credit for qualifying investments in AI security research and responsible development
- Designed to complement existing R&D incentives while specifically targeting security initiatives

### 2. *Qualification Requirements*

- Investments must exceed previous year's baseline spending on AI security research (maintenance of effort provision)
- Research findings must be published to benefit the broader AI development community (subject to appropriate security considerations and redaction protocols)
- Developers must maintain and update a public Responsible Scaling Policy (RSP)

### 3. *Scope of Qualifying Research*

The framework encompasses eight key areas of AI security and responsible development:

- *AI Governance*: Development of oversight frameworks and risk management practices
- *AI Security*: Development of protections against adversarial attacks and

---

<sup>1</sup> In a viral [resignation post](#), one of them stated: “Sometimes we were struggling for compute and it was getting harder and harder to get this crucial research done . . . safety culture and processes have taken a backseat to shiny products.”

- unauthorized access
- *AI Safety*: Investigation of robustness and fail-safe mechanisms
- *Bias and Fairness*: Methods for identifying and mitigating discriminatory outcomes
- *AI Alignment*: Approaches to ensuring AI systems align with human values
- *Transparency*: Advancement of interpretability and explainability measures
- *Data Privacy*: Development of privacy-preserving methodologies
- *Human-AI Collaboration*: Research on effective human-AI interaction paradigms

## Implementation Framework

The Treasury Department and IRS will administer the credit and may consult with NIST, the FTC, and other relevant agencies in setting definitions and eligibility requirements.

Each participating organization must publish and maintain an RSP that includes the following core elements:

- Risk assessment frameworks
- Ethical development commitments
- Security protocols
- Bias mitigation strategies
- Transparency measures
- Incident response procedures
- Human oversight mechanisms

To ensure both a sufficient timeline for investment planning and encourage a review of the credit's effectiveness in light of technological developments, we recommend including a ten year sunset clause.

## Anticipated Impact

This tax incentive structure aims to create a self-reinforcing cycle of secure and responsible AI advancement. By reducing the net cost of safety and security research, we can help ensure that advances in AI capabilities are matched by corresponding advances in safety and security measures.

The public disclosure requirements will facilitate the development of an AI security research ecosystem while the RSP requirement will provide the research community with concrete metrics for evaluating developers' commitment to responsible development. Through this balanced approach, we can help ensure that the remarkable potential of AI technology is realized while maintaining strong security standards and protecting the public from harm.