As of: 5/29/25, 6:31 AM
Received: May 28, 2025

PUBLIC SUBMISSION

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0135
Comment on FR Doc # 2025-07332

Submitter Information

Name: Theo Curtis

General Comment

Response to the 2025 National AI R&D Strategic Plan Request for Information
Submitted by Theo Curtis

Docket ID No. NSF-2025-OGC-0001
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 2025 National AI R&D Strategic Plan and associated documents without attribution.

Title: Scaling AI Safety Research to Secure Long-Term U.S. Leadership

To sustain its leadership in artificial intelligence, the United States must pair rapid innovation with proportionate investment in AI safety research. As AI systems become more capable — particularly large-scale, general-purpose models, they also become more difficult to audit, predict, and align with human values. Robust and forward-looking investment in AI safety, especially in areas such as red-teaming, interpretability, and alignment, is essential to ensure these technologies serve public interests over the long term.

Priority Research Areas for AI Safety (2025–2030)
We recommend that the 2025 Strategic Plan explicitly prioritize foundational AI safety challenges that are unlikely to be addressed through market incentives alone, including:

1. Scalable Red-Teaming and Evaluation Frameworks
Red-teaming - adversarial testing of AI systems to uncover hidden risks and failure modes — must become a first-class research priority. Current red-teaming efforts are ad hoc and under-resourced. A strategic federal investment would enable:

Standardized safety benchmarks for capabilities such as deception, manipulation, or unauthorized tool use.

Stress testing under diverse threat models, including dual-use concerns, misuse by malicious actors, and autonomous goal-seeking behavior.

-Independent evaluation infrastructure, potentially coordinated through national labs or federally funded research centers, to test the safety properties of proprietary frontier models.

-Red-teaming at scale, including synthetic environments and simulations that expose edge-case behaviors and unintended capabilities.

-Red-teaming is one of the most practical levers to identify catastrophic risks before deployment. Federal coordination would also support international norms for responsible AI evaluation.

2. Interpretability and Mechanistic Transparency
As models grow in complexity, so does the need to understand how and why they behave the way they do. Interpretability research aims to make opaque systems legible to humans — a foundational requirement for meaningful oversight. Strategic priorities should include:

-Mechanistic interpretability: Methods that reverse-engineer the internal structure of large models, identifying circuits, representations, and submodules responsible for specific behaviors.

-Behavioral interpretability: Tools to explain model outputs under specific inputs or contexts, enabling better understanding of reasoning chains or latent knowledge.

-Scalable oversight tools: Human-in-the-loop systems to audit or guide model reasoning during training and deployment, especially when human supervision is partial or limited.

-Interdisciplinary collaborations: Support for joint efforts between AI researchers, cognitive scientists, and security experts to ensure interpretability tools meet real-world needs.

These efforts are essential for building verifiable, steerable AI systems that are robust under novel or adversarial conditions.

3. Alignment and Corrigibility Research
A central open problem in AI safety is value alignment — designing systems that pursue goals consistent with human intent, even in complex, evolving, or ambiguous contexts. Key research directions include:

Robust reward modeling: Training methods that infer or learn human preferences across diverse domains while avoiding reward hacking or specification gaming.

-Safe exploration and goal generalization: Techniques to ensure that AI systems generalize intended behavior to unfamiliar environments without unsafe trial-and-error.

-Corrigibility: Designing agents that remain amenable to human intervention, including stopping, redirecting, or updating their goals — even when acting autonomously.

-Uncertainty-aware systems: Ensuring that AI systems can recognize when they are out of distribution or uncertain, and defer decisions or escalate to human review.

Alignment is not only a theoretical concern. As models are given more autonomy — in military, scientific, or economic domains — their ability to reason safely under uncertainty becomes critical to national security and economic stability.