AI Dominance Requires Interpretability and Standards for Transparency and Security

David Bau, Tom McGrath, Sarah Schwettmann, Dylan Hadfield-Menell *

March 2025

Executive Summary

Dominance in AI will come from achieving *interpretability*. In past technological revolutions such as biotechnology, interpretability has been the key to mastery. Although AI interpretability will be more difficult than the creation of AI, research has shown that interpretability is both achievable and essential. Innovation in AI interpretability depends on computational transparency, but unfortunately, the American AI ecosystem lags behind foreign competitors by blocking the access required for interpretability. Systems such as the National Deep Inference Fabric (NDIF) provide secure computational transparency without parameter copying. To lead, the U.S. needs to establish uniform computational transparency in AI.

We recommend:

- **Provide sustained funding for interpretability research initiatives such as NDIF.**
- **Establish an AI Interpretability and Control Standards Working Group within NIST.**
- **Direct NSF, DOE, and DOD to build and allocate dedicated computational resources for interpretability research.**

Authors

David Bau (PhD MIT, MS Cornell, AB Harvard) is Assistant Professor of Computer Science at Northeastern University, Director of the National Deep Inference Fabric, and a leading expert on AI interpretability. His lab develops methods that allow scientists to make sense of the calculations within AI. Prior to his academic work, Prof. Bau developed widely-used products in industry including search algorithms at Google and web browsers at Microsoft.

*Correspondence: davidbau@northeastern.edu, tom@goodfire.ai, sarah@transluce.org, dhm@csail.mit.edu. 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.

Tom McGrath (PhD Imperial College, MMathPhys Warwick) is Chief Scientist and co-founder at Goodfire AI, a leading US interpretability research startup. Goodfire AI develops and applies mechanistic interpretability techniques to advance our ability to understand, edit, and control AI, and has partnered with leading US biomedical organisations to apply these techniques. Prior to his work at Goodfire, Dr. McGrath was a researcher at Google DeepMind, where he worked on developing and understanding frontier language models and the AlphaZero agent.

Sarah Schwettmann (PhD MIT, BS Rice) is Chief Scientist and co-founder of Transluce, a non-profit research lab working toward responsible development and deployment of AI in the public interest. Transluce builds AI-backed tools for automatically understanding the representations and behaviors of AI systems, and contracts with labs and governments to audit frontier AI systems for security risks, surprising behaviors, and novel capabilities. Sarah is also a Research Scientist in MIT's Computer Science and AI Laboratory, where her research group has developed some of the first large-scale AI-backed interpretability pipelines.

Dylan Hadfield-Menell (PhD UC Berkeley, MEng and SB MIT) is Assistant Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT). He leads the Algorithmic Alignment Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL), focusing on developing methods to ensure that AI systems' behavior aligns with the goals and values of their human users and society as a whole. His research seeks to enable safe and effective human-AI interaction and support meaningful human control of AI systems. He has been recognized as an AI2050 Early Career Fellow by Schmidt Futures.

Interpretability is the Key to AI Leadership

Currently, AI systems are black boxes; many ambitious AI applications cannot be built, because of the field's inability to understand and control AI mechanisms. Existing AI systems cannot be effectively implemented because they suffer from unpredictable weaknesses and errors. To lead in AI requires mastery of the internal calculations.

The role of AI interpretability is analogous to the role of biochemistry in medicine. The dominant companies in biology and medicine do not blindly breed new species or guess new medicines; their mastery comes from detailed understanding of genes and chemistry. As a result, modern biology is dominated by biochemistry—biological interpretability—rather than just breeding.

Interpretability rather than mere access has also been the key for dominance on the Internet. Twenty-five years ago, the invincible technology company was AOL: it controlled Internet access for millions of users, was valued at $350 billion (like OpenAI today), and seemed to have a stranglehold over the industry. In retrospect it is obvious why AOL fell from glory. It was built on the erroneous assumption that controlling Internet access would be the key to Internet dominance. They did not offer any serious solutions to "Internet interpretability."

Instead, the Internet has been dominated by the companies that harness the complexity of the open web to make it useful and understandable to humans: from Google to Amazon to Meta, the winning companies are all masters at the art of making the web *interpretable to people*. They are leaders at collecting, organizing, understanding, recommending and explaining torrents of Internet data, distilling human understanding and value from the chaos of content.

Unfortunately, the U.S. AI industry is caught in the same trap that brought down AOL. Our major AI companies are attempting to create a closed-AI world with the mistaken idea that AI dominance will come from training and controlling access to large-scale AI models whose internals remain deeply mysterious to users and experts. This closed-AI model is as flawed as AOL.

As AI develops superhuman capabilities, the industry will be dominated by the future companies that clarify the mystery in AI and make it useful and understandable to humans. The main challenge will be to make the new knowledge in AI *interpretable to people*. We will need to become leaders at the "biochemistry of AI", that is, collecting, organizing, understanding, recommending and explaining knowledge from the massive complexity of AI calculations, distilling human understanding and value from the tangle of neural network connections.

Because achieving AI interpretability will require years more innovation than just training AI, for the U.S. to maintain its dominance in AI, we must incubate a dynamic industry in which upstart innovators are empowered to address the AI interpretability problem in the long run.

Interpretability Bridges the Human-AI Knowledge Gap

The most valuable aspect of AI will be its *knowledge beyond human knowledge*. By definition, the knowledge contained within AI that humans do not yet know will not be planned or evaluated ahead of time, and will require intrpretability methods to unlock. This AI knowledge may be either useful—such as a clever new way to solve a problem—or unwanted—such as a tricky new way to deceive the user.

While some AI experts worry that the emergence of superhuman AI knowledge will pose an intractable problem for humanity—AI pioneer Geoff Hinton explains that, when humans face superhuman AI, "we'll be the three-year-olds"—the authors are experts in the field of AI interpretability, and we can report that human understanding of superhuman AI is both achievable and essential. The key is the computational transparency of AI: unlike the impossibility of outwitting a smarter human opponent, we can always crack open AI and inspect its internal calculations. Computational transparency means, with the right tools, no cognitive mystery is beyond reach.

Interpretability is necessary for powerful applications of AI. Scientific superintelligences—AI systems with superhuman scientific knowledge—are already in use today. For example, AlphaFold and Evo 2 predict key parts of biological systems better than humans have ever been able to do, and AlphaZero is superhuman at the games of Go and chess. Understanding these superintelligences is key to scientific discovery with AI; a key element of the Executive Order. Although their superhuman knowledge may seem inscrutable to people, the information is locked up inside their internal calculations, waiting for interpretability tools to set it free.

The means to extract knowledge from AI are within reach: recent breakthroughs in interpretability have allowed researchers to extract new concepts from AlphaZero to teach top-level chess Grandmasters new concepts (with one of these players going on to become World Champion). Similarly, scientists have begun to understand concepts inside the state-of-the-art biology model Evo 2, extracting tens of thousands of features which are being analyzed for new scientific insights in the field of genomics. The techniques that have made this possible are in the early stage of development, and need to be supported and developed to achieve their full potential.

In large language model interpretability research, the current frontier is the challenge of understanding "reasoning" language models that have been trained to perform deductions using a long internal monologue. Preliminary research reveals neural fingerprints of iterative search processes, suggesting the presence of unspoken internal search goals.

Unfortunately for the U.S. AI ecosystem, research progress in reasoning-model interpretability is focused on the Chinese DeepSeek R1 model, despite the superiority of OpenAI's reasoning models that were invented in the U.S. and deployed earlier. The research focus on DeepSeek R1 arises because OpenAI has not provided any computational transparency. Today, Chinese reasoning models are the only ones that provide the technical prerequisites for interpretability research. **This unfortunate situation puts Chinese AI, for the first time, in the leading position in the latest work in AI interpretability.**

The U.S. Lags in Technical Prerequisites for Interpretability

The key to AI interpretability is computational transparency. No matter how complex the AI, we can crack it open and inspect its internal calculations, which means that with the right tools, no cognitive mystery is beyond the reach of human understanding. Unfortunately, the American approach is closed interfaces that *do not provide computational transparency*, and this is strangling progress.

In contrast, China appears to be building its AI community around an open-model-parameter consensus that does provide computational transparency. If this imbalance persists, then the Chinese open ecosystem will beat the closed American establishment; their dynamic community will enable innovations in AI that ours does not.

The US AI marketplace has only one company, Meta, that stands alone in releasing large models openly. The half dozen other major US AI providers have failed to adopt this approach, and as a result, the US national AI industry is fragmented and disorganized. Entrepreneurs can not build freely on computational transparency, because the uncertainty of access forestalls major investments in scalable AI interpretability and gives away an advantage to AI copycats and foreign competitors.

To create an American AI ecosystem that provides uniform computational transparency while also providing security requires coordination: we need a well-designed technical standard for *transparent and secure* AI access.

A Standard for Secure Computational Transparency

The National Deep Inference Fabric (NDIF) demonstrates a technical path for providing computational transparency without enabling copycats. It allows model providers to retain control over their own parameters, preventing exfiltration, while allowing customizers to freely innovate within the computations of AI inference by running complex customization code within the fabric.

As shown in Figure 1, the partial openness of NDIF is analogous to the partial openness of the Internet. On the open Internet (1b), code on the server remains a secret while the code sent to

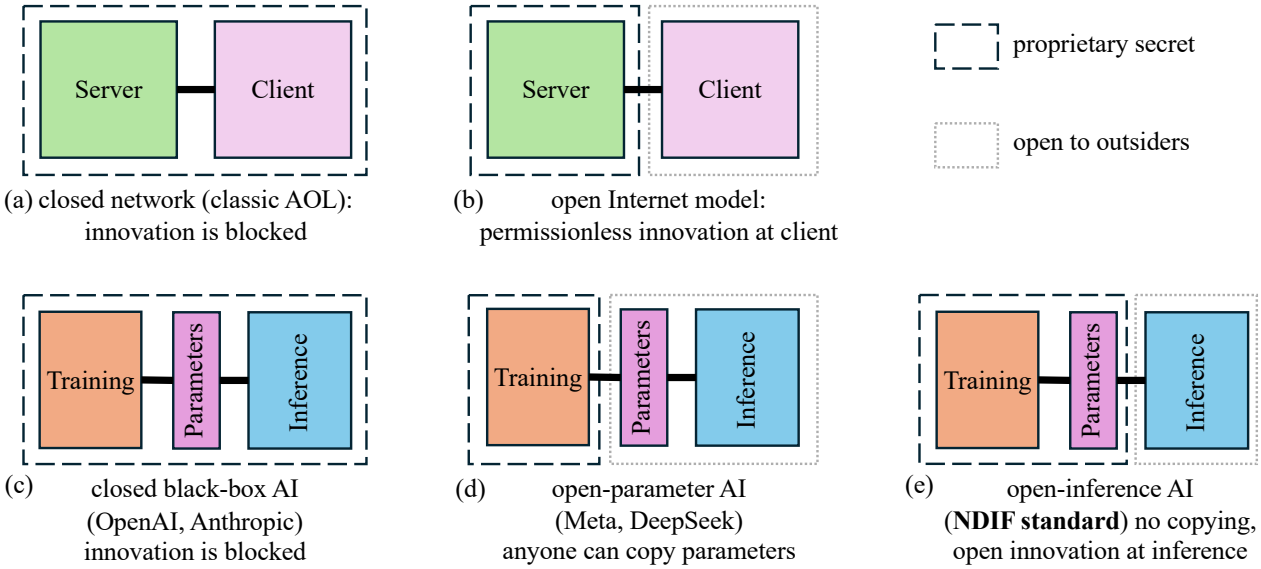

Figure 1: Forms of closed and open access on the Internet and in AI. Both (a) closed networks and (c) closed black-box AI block innovation. (b) The open internet is analogous to (d) open-parameter AI; both enable permissionless innovation on the outputs of proprietary serving or training, but they also allow copying. (e) NDIF open-inference AI does not open parameters, so it precludes copying while enabling innovation in inference.

the client becomes freely visible, allowing essential information to be analyzed and organized by third parties. In Meta's open-parameter model (1d), the training details remain private while the entire inference process including parameters are made public, allowing innovation by third parties and also encouraging the parameters to be copied. In the NDIF standard (1e), training details and parameters remain private, precluding copying, while the inference calculations becomes public inside the fabric, enabling research and innovation.

NDIF achieves this partial privacy by defining a standard for inference-customization and analysis code to be transported and executed within the same secure fabric as the AI parameters. This kind of interface provides the computational transparency needed for AI interpretability research and development, while allowing AI providers to monitor use and limit download bandwidth. When combined with network security and monitoring, this access model can enable innovation while minimizing risk of copying or exfiltration of model parameters.

Since the NDIF approach requires researchers to do their work within a secure fabric, an ecosystem built around such a standard will need to provide other prerequisites for permissionless innovation: secure computational resources sufficient for entrepreneurs and researchers to use NDIF, and a stable and neural access structure that protects businesses who wish to build a scalable business within the secure fabric.

Combined transparency and security needs to become the U.S. AI standard. This will allow for the emergence of rapid innovation while preventing unrestrained copies of our most powerful AI models.

Conclusion and Recommendations

We recommend:

- **Provide sustained funding for interpretability research initiatives** such as the National Deep Inference Fabric (NDIF). Existing initiatives—including NSF's Directorate for Technology, Innovation and Partnerships, DARPA's AI Forward program, DOE's Advanced Scientific Computing Research (ASCR) program, and the NITRD AI R&D Interagency Working Group—should make funding and coordinating interpretability research a national priority. This funding should support core infrastructure development such as NDIF, as well as grants to academic and private sector researchers pursuing novel interpretability techniques.
- **Establish an AI Interpretability and Control Standards Working Group within NIST** to develop technical standards for computational transparency and model security. This working group should codify best practices, interoperable standards, and research priorities for interpretability research at scale.
- **Direct NSF, DOE, and DOD to build and allocate dedicated computational resources for interpretability research.** These resources should be made available to qualified researchers through streamlined access mechanisms, with priority given to projects focused on understanding the internal mechanisms of frontier AI systems.

America is leading at a pivotal moment in the development of AI. But as AI systems begin to surpass human knowledge, the most important progress in the field will turn from training to interpretability. To maintain American dominance in this next phase of AI, our country needs to adopt an AI access standard that provides the computational transparency to enable free innovation in interpretability. NDIF shows how such a standard is possible while maintaining security of large AI model parameters.