

# Disincentivizing Bioweapons

• Theory & Policy •  
• Approaches •



Edited by Nathan A. Paxton

NTI:bio

## ⋮ Acknowledgments

The editor and authors gratefully acknowledge the support of those who were instrumental in the development of this report, including the many expert contributors who generously shared their time and expertise. We would like to thank NTI's Communications team—in particular, Scott Nolan Smith and Mimi Hall—for their invaluable assistance with editing and production and NTI | bio Program Officer Gabby Essix for her assistance with editing and managing this project. We are deeply grateful to Longview Philanthropy for supporting this work.

© 2024 Nuclear Threat Initiative



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

The views and opinions expressed in this collection of essays are those of the individual authors and do not necessarily reflect the official policy or position of the Nuclear Threat Initiative (NTI), its Board of Directors, or any of the institutions with which the authors are affiliated.

The inclusion of any essay in this collection does not imply endorsement by NTI of the opinions expressed within.

## ⋮ About the Nuclear Threat Initiative

The Nuclear Threat Initiative is a nonprofit, nonpartisan global security organization focused on reducing nuclear, biological, and emerging technology threats imperiling humanity.



# Disincentivizing Bioweapons

• Theory & Policy •  
• Approaches •



# Contents

Introduction .....	1
Section 1: A Tactical Framework to Shape Intention and Disincentivize State Biological Weapon Development and Use .....	9
Guarding Against Catastrophic Biological Risks: Preventing State Biological Weapon Development and Use by Shaping Intentions .....	11
The Role and Limits of Transparency Measures in Disincentivizing Biological Weapons .....	33
Attribution as Deterrence for Biological Weapons .....	49
After Bioweapons—What? Accountability for Development and Use of Biological Weapons.....	63
Section 2: Disincentivization Challenges That Require Further Attention.....	77
Two Competing Bioweapons Nonproliferation Policies: Deterrence by Denial and Dissuasion .....	79
The Biological Weapons Taboo: A “New” Focus for Arms Control ....	103
Prospects for Assessing State Intent to Proliferate Biological Weapons .....	115
Biotechnology and the Dead Zone for Managing Dual-Use Dilemmas.....	133
“Emergent Abilities,” AI, and Biosecurity: Conceptual Ambiguity, Stability, and Policy.....	149
Section 3: Potential Tools and Narratives for Dissuasion and Deterrence.....	165
Simple Tool for Disincentivizing the Worst Pandemic Bioweapons....	167
Appendix .....	181
Meeting Report: November 2023 Workshop on Disincentivizing State Bioweapons Development and Use.....	183

# “Emergent Abilities,” AI, and Biosecurity: Conceptual Ambiguity, Stability, and Policy

Alex John London

## Summary

Recent claims that artificial intelligence (AI) systems demonstrate “emergent abilities” have fueled excitement but also fear grounded in the prospect that such systems may enable a wider range of parties to make unprecedented advances in areas that include the development of chemical or biological weapons. Ambiguity surrounding the term “emergent abilities” has added avoidable uncertainty to a topic that has the potential to destabilize the strategic landscape, including the perception of key parties about the viability of nonproliferation efforts. To avert these problems in the future, scientists, developers, policymakers, and other parties should take credible steps to strengthen the health of the scientific ecosystem around AI.

## Introduction

Recent advances in AI, specifically generative AI, which includes generative pretrained models or large language models (LLMs), have captured the public imagination and set off alarm bells among the many parties interested in security. At the epicenter of this concern are claims that with increases in the scale of compute, volume of training data, and number of parameters, predictable gains in performance<sup>1</sup> have been accompanied by powerful and surprising, emergent abilities.<sup>2</sup> These range from the ability to plan,<sup>3</sup> to reason about causal relationships,<sup>4</sup> and most



surprising of all, to demonstrate “sparks of [artificial general intelligence] AGI,” or early signs that these systems are on the verge of constituting artificial general intelligence.<sup>5</sup> At the extreme, such claims conjure fears that generative AI will turn on humanity like *The Terminator’s* Skynet, *War Games’s* WOPR, *2001: A Space Odyssey’s* HAL-9000, or Marvel Comics’ Ultron. But they also engender slightly less fanciful worries that powerful new capabilities might enable a wider range of players, from rogue states or malevolent organizations to highly motivated individuals, to more easily, quickly, or cheaply develop biological weapons, including new agents with enhanced lethality.<sup>6</sup>

If technological advances that are relatively easy to access truly can produce revolutionary new threat capabilities for a wider array of parties, strategic equilibria can be destabilizing. If actors believe that they can strengthen their position by acquiring these new capabilities, efforts at nonproliferation can be undermined. Even if it is not clear that such technological advances have materialized, sufficiently credible uncertainty about technologically assisted threat capability can create a destabilizing environment in which actors feel compelled to act, either to strengthen their strategic position or to mitigate risks that might compromise their current position. As a result, and as illustrated below, uncertainty about the capabilities of new AI systems can reach beyond commercial interests to impact the larger strategic landscape—the way actors represent the basic features that frame decision problems related to security.

Conceptual ambiguities have exacerbated the challenge of crafting evidence-based policy, and those ambiguities have sown confusion, obscured the nature of stakeholder disagreements, and fostered an atmosphere of hype. To better navigate such challenges—including the potential development of bioweapons through the assistance of AI systems—key stakeholders should take steps to strengthen the health of the scientific ecosystem surrounding AI.



## Conceptual Ambiguities around "Emergence" and "Ability"

The one ability that LLMs do possess is the one for which they have been designed and trained—to predict the next token given a set of tokens presented in the form of a prompt. (A token is a set of letters, often a pair or a triple.) It is genuinely stunning that systems trained to build complex statistical relationships among tokens in incomprehensibly large training sets can produce coherent text that is often relevant to the prompt and sometimes surprisingly useful. This facility with language has led researchers to inquire about what other abilities these systems might possess. It is in this context that researchers have claimed to identify emergent abilities. Unfortunately, both terms in this phrase are ambiguous—and this ambiguity has important implications for risk and for judgments of safety and reliability.

Consider first what might be meant by something being emergent or emerging at some level of complexity. “Emergent” may have two distinct meanings here. Epistemic matters relate to the nature of knowledge and how knowledge is validated. The *epistemic sense* of emergent refers to the difficulty in predicting, at one level of complexity, what a system might be able to do at some higher level of complexity.<sup>7</sup> Ontology refers to the nature of being or existence. The *ontological sense* of emergent refers to something new coming into existence, to the birth of a new ability.<sup>8</sup>

Now consider what might be meant by some new “ability.” This term might refer to the *task* that a system can be used to perform or to the internal *capacities* by virtue of which it is able to perform some tasks. The disambiguated combinations of these views (two meanings of “emergent” and two of “ability”) are summarized in Table 1 and numbered for ease of reference.

To understand the extent of the ambiguity in these terms, consider now the extent of diversity in the disambiguated views this phrase can express.



TABLE 1: AMBIGUITIES IN THE CONCEPT OF “EMERGENT ABILITIES”

		Ambiguities in the notion of ability	
		Task related	Capacity related
Ambiguities in the notion of emergence	Epistemic	<p>1. The mundane claim that as scale increases, one may not know how model performance will increase on new tasks.</p>	<p>3. The deeper uncertainty about whether increases in scale will result in models with surprising new internal capabilities (uncertainty about whether models will develop unexpected capacities that will enable them to perform tasks that humans currently cannot).</p>
	Ontological	<p>2. The more mundane claim that internal capacities remain the same but the surprising claim that predicting the next token can be a useful approach to performing a much wider range of tasks than initially thought.</p>	<p>4. The amazing claim that, at some new scale, systems develop new internal capacities in virtue of which they can better perform established tasks or perform a wider range of new tasks. Necessary for artificial general intelligence and a presupposition of many surprising claims about internal representations.</p>





## ⋮ Dangerous Ambiguity: Evolution or Revolution?

Expressions that fall into Boxes 2 and 4 in Table 1 make statements about how to understand what generative AI can do and how it does that. To use slogans, Box 2 sees generative AI as a largely evolutionary progress, while Box 4 treats it as a revolutionary leap.

When talk of emergent abilities falls into Box 2, it asserts the claim that with increases in scale and complexity, a system that predicts the next token can be used to perform a wider range of tasks than simply producing or predicting the next token. These assertions are likely to be somewhat measured when it comes to claims about the reliability or robustness of these systems since outputs are generated from complex statistical relationships among tokens without the assertion that such systems are learning the underlying structure of some domain or developing some novel cognitive ability.

Figure 1A illustrates the underlying process that generates such results. Those who hold this view are likely to regard these systems as “stochastic parrots.”<sup>9</sup> The utility of a stochastic parrot depends on preserving correlations among syntactic relationships derived from the data fed into the model that those models use to associate inputs with correct or useful outputs. From this perspective, the capabilities of these systems, while impressive, are limited to compressing and making available information that is already contained in the training data—they repackage the portions of the Internet on which they have been trained.<sup>10</sup> A reasonable expectation in the possible state of the world described by Box 2 in Table 1 and Figure 1A is that confabulations (so-called hallucinations) are likely to be endemic to such systems since they combine tokens without tracking the underlying logical or causal structure of the world.<sup>11</sup>

In contrast, Box 4 in Table 1 contains assertions that, at some new scale, LLMs develop new internal capabilities—the ability to reason and plan, for example—in virtue of which they can perform better on established tasks or perform a much wider range of tasks. Such views are illustrated in Figure

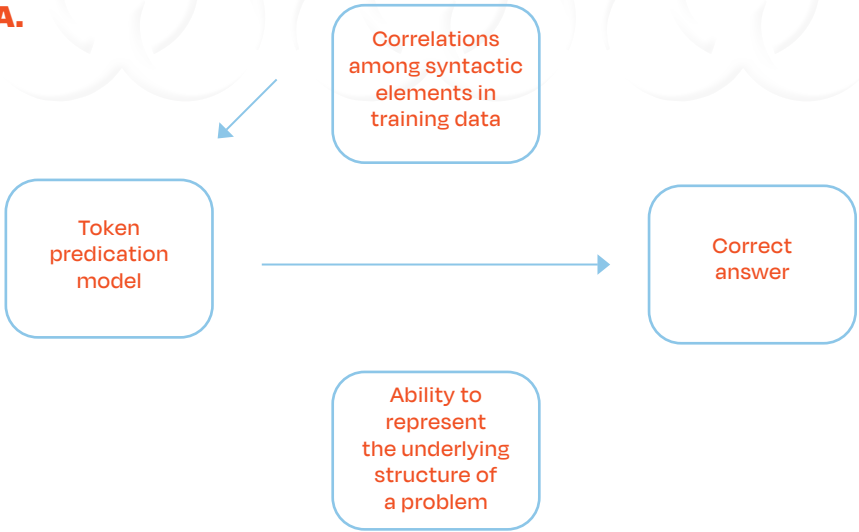


1B. Some of the most sensational claims about LLMs reside in this box because they imply that designing a system to predict the next token can, with sufficient complexity, give rise to internal representations that constitute a revolutionary leap in cognition. If such systems form an internal representation of the structure of some domain—such as biochemistry—that allows them to reason and plan, then they might be able to solve problems that go far beyond the simple application of prior patterns derived from the training data. This includes discovering new cures or new toxins or pathogens that are currently beyond human reach. Similarly, the hope—or concern, if an actor’s motives are malign—is that by tuning these models to rely more heavily on new capabilities, fabrications or hallucinations might be eliminated, and these systems might become more reliable and robust.

## ⋮ Risk, Uncertainty, and the Strategic Landscape

**A**mbiguities about the claim being expressed by the phrase “emergent abilities” have major implications for how to think about the strategic landscape. These features include the states of the world that actors regard as feasible—all the things that might happen, from natural disasters and pandemics to terrorist attacks and other acts of aggression (and whether this should include the novel actions of a new AI). They also include the set of acts actors might take to avert, mitigate, or respond to various threats.

**A.**



**B.**

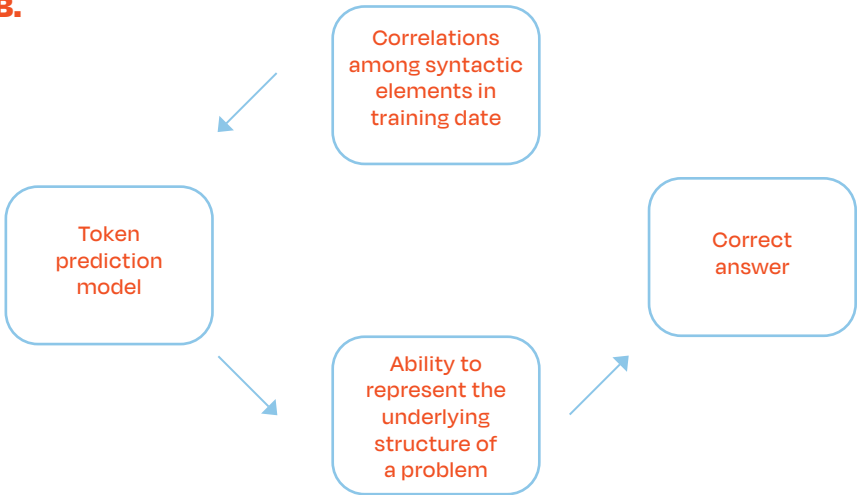


Figure 1: Directed acyclic graphs (DAGs) representing the difference between (A) models that produce the correct answer merely through correlations between syntactic elements in their training data and syntactic features of the correct answer and (B) models that develop new internal representations that allow them to produce the correct answer by exercising some new capability. DAG A corresponds to Box 2 in Table 1; DAG B corresponds to Box 4 in Table 1.



When talk of emergent abilities falls within Box 2 in Table 1/Figure 1A, it represents evolutionary progress. It may pose practical challenges for interested parties, but the uncertainty it generates falls into Box 1 in Table 1. Managing this kind of uncertainty constitutes a problem of decision-making under risk because the states of the world that are feasible and the set of acts that are available are largely stable. This is the kind of uncertainty with which strategic actors must routinely grapple. In contrast, developments captured by Box 4 in Table 1/Figure 1B entail the kind of uncertainty expressed in Box 2. This uncertainty represents a substantive alteration to the strategic landscape. The reason is that systems that develop new internal capacities might be able to do things that are qualitatively new. In particular, emergent abilities of this type might not simply enable systems to do things that had been envisioned by being assigned a low probability (the kind of uncertainty represented in Box 1).

Instead, they might enable abilities that are unexpected in the sense that actors had not thought to consider them when envisioning the states of the world that might arise or when enumerating the set of acts they should consider. This is sometimes called “Knightian uncertainty,” where the decision-maker is unsure about which states of the world to entertain, let alone what probability to assign to them.<sup>12</sup> For example, is there a need to contemplate and prepare for events in which systems with new internal capacities develop biological weapons of a type that humans have yet to envision and use some novel mechanism to deliver them in a way that produces an outcome not contemplated? Without a well-defined partition over qualitatively new and unexpected events, it is difficult to assign coherent probabilities to each state.

A common reaction to uncertainty of this kind is to move away from standard approaches to decision-making under risk, in which potential harms and benefits are multiplied by their probability of occurring, and permissible acts are those that produce a ratio of expected benefit and harm that is “reasonable” in some sense. Instead, with Knightian uncertainty, some actors will gravitate toward approaches that are more precautionary and loss averse in that they give priority to averting outcomes that would be extremely bad, no matter how likely those outcomes are to occur.<sup>13</sup> When actors are unsure about the type of challenge they face—whether updating prior assessment or having to



imagine wholly different worlds—these ambiguities can change not only how stakeholders think about novel technologies and their possible impacts but the decision rule they use to reason about risk and uncertainty—and thus the trade-offs between security and liberty that they view as reasonable.<sup>14</sup>

## ⋮ Equivocation, Instability, and Policy

The difficulty of formulating coherent and ethically sound security policy is exacerbated when implicit conceptual confusion leads different agents to radically different representations of the strategic landscape. The realistic possibility that parties who see LLMs as capable of supercharging chemical or biological weapons programs will view nonproliferation efforts as infeasible or view states that develop LLMs as violating prohibitions on chemical or biological weapons programs, which can have a destabilizing effect. As a consequence, states looking to reinforce nonproliferation efforts may contemplate, among other steps, restrictions on AI work that violate the rights and liberties of individuals or groups.

A healthy scientific ecosystem is a bulwark against such uncertainty. The health of the scientific ecosystem is facilitated by three elements. The first is drawing clear conceptual lines between unambiguous views that are well-differentiated alternatives. The goal here is to ensure that the various properties of systems associated with each view can be carefully articulated so that various claims about utility and hypotheses about the emergence of novel capacities can be differentiated. An ecosystem in which ambiguous claims frustrate the ability of interested parties to efficiently differentiate relevant alternatives in terms that can be empirically tested is unhealthy.

Second is a process that promotes rigorous, expeditious, and efficient testing designed to identify which of these claims are supported by evidence. Efforts to evaluate systems under conditions that control for confounding, and thereby that distinguish between the states of affairs depicted in Figure



1A versus B, should be central to this work. Third is a credible system of incentives that reward engaging in this process of rigorous testing and peer evaluation before claims about the abilities of systems are widely publicized. Central to this process should be credible efforts to reduce conflicts of interest that arise when the parties who profit from the sale of a system are also producing the research that outlines system capabilities, potential benefits, and shortcomings.

In contrast, early studies that purport to substantiate claims from Box 4 without carefully distinguishing and controlling for mechanisms for model performance that fall within Box 2 exacerbate uncertainty and perpetuate the prospect of inflated expectations. Once public attention has been captured and expectations framed, the buzz created by inflated claims, whether of benefit or danger, overshadows and threatens to drown out the more measured findings of carefully controlled investigations. As one example, the claim that LLMs have developed the capacity to plan is central to the hopes of AGI optimists and the fears of AGI pessimists who trace out alternatively utopian and dystopian visions of the future. In a series of studies, Subbarao Kambhampati and colleagues<sup>15</sup> evaluated the planning capabilities of generative AI models in a way that differentiates task performance based on recall or pattern recognition between syntactic elements in the prompt and syntactic elements in the training data (Box 2 and Figure 1A) from the capacity to represent and reason about the underlying structure of a planning problem (Box 4 and Figure 1B). When the syntactic elements used to refer to items in a planning problem are altered, but the structure of the problem remains unchanged, LLM performance effectively disappears. Similar findings have been reported in studies that examine causal reasoning,<sup>16</sup> theory of mind,<sup>17</sup> and the more general claim that novel capabilities emerge suddenly at new scales.<sup>18</sup>

As a result, it is unsurprising that early reports that generative AI models might be used to develop novel chemical or biological agents have been tempered by recent findings that such models offer marginal advantages when compared to the baseline of using information already present on the Internet.<sup>19</sup> Improving



the health of the scientific ecosystem around AI cannot eliminate uncertainty that attends new scientific advances, but it can help stakeholders reduce avoidable uncertainty that arises from conceptual ambiguity.

Strengthening the health of the scientific ecosystem surrounding AI is critical for ensuring that research in this area produces information on which a wide range of decision-makers can rely when making decisions that can impact the rights and well-being of large numbers of people.<sup>20</sup> Because Knightian uncertainty can destabilize strategic equilibria, practices that minimize the perception of such uncertainty when this perception is avoidable help to avert circumstances in which actors might feel compelled to defect from nonproliferation efforts. They also help ensure that public perception, stakeholder attention, social resources, and security efforts are not captured by parties who might benefit from inflated perceptions regarding the abilities of novel technologies.

## ∴ Conclusion

**T**he challenge of balancing security with the freedoms that define open societies is complicated by advances in technology. These complications stem from uncertainty around the disruptions that will flow from innovation as well as the challenges that new approaches can pose to old concepts. Conceptual clarity is essential to the ability of stakeholders in the scientific ecosystem to expeditiously articulate and efficiently address pivotal scientific questions, to maintain a realistic sense of the strategic landscape, to mitigate the dangers of hype, and to foster the creation of timely, evidence-based policy.



## ⋮ About the Author

### **Alex John London**

*K&L Gates Professor of Ethics and Computational Technologies and  
Director of the Center for Ethics and Policy, Carnegie Mellon University*

Alex John London, PhD, is the K&L Gates Professor of Ethics and Computational Technologies at Carnegie Mellon University. His book, *For the Common Good: Philosophical Foundations of Research Ethics*, is available in hard copy from Oxford University Press and electronically as an open access title. He is currently a member of the U.S. National Science Advisory Board for Biosecurity, and he has served as an ethics expert in consultation with numerous national and international organizations, including the World Health Organization, the World Medical Association, the U.S. National Academies, and the U.S. National Institutes of Health.

## ⋮ Acknowledgments

I thank Jonathan Herington, Peter Spirtes, and Subbarao Kambhampati for critical feedback on an early draft of this manuscript.





## Endnotes

- <sup>1</sup> Jordan Hoffmann et al., “Training Compute-Optimal Large Language Models,” [arXiv:2203.15556](https://arxiv.org/abs/2203.15556).
- <sup>2</sup> Jason Wei et al., “Emergent Abilities of Large Language Models,” [arXiv:2206.07682](https://arxiv.org/abs/2206.07682).
- <sup>3</sup> Wenlong Huang et al., “Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents,” *PMLR* 162 (2022): 9118–47.
- <sup>4</sup> Emre Kıcıman et al., “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality,” [arXiv:2305.00050](https://arxiv.org/abs/2305.00050).
- <sup>5</sup> Sébastien Bubeck et al., “Sparks of Artificial General Intelligence: Early Experiments with GPT-4,” [arXiv:2303.12712](https://arxiv.org/abs/2303.12712).
- <sup>6</sup> Matthew E. Walsh, “Why AI for Biological Design Should Be Regulated Differently Than Chatbots,” *Bulletin of the Atomic Scientists*, September 1, 2023, [thebulletin.org/2023/09/why-ai-for-biological-design-should-be-regulated-differently-than-chatbots](https://thebulletin.org/2023/09/why-ai-for-biological-design-should-be-regulated-differently-than-chatbots); Alexei Grinbaum and Laurynas Adomaitis, “Dual Use Concerns of Generative AI and Large Language Models,” *Journal of Responsible Innovation* 11, no. 1 (2024): 2304381.
- <sup>7</sup> For example, Ganguli and colleagues use the epistemic sense when they note the “high unpredictability” of large generative models in the sense that “specific model capabilities, inputs, and outputs can’t be predicted ahead of time” (Deep Ganguli et al., “Predictability and Surprise in Large Generative Models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* [2022]: 1747–64).
- <sup>8</sup> Wei and colleagues take the ontological sense of emergence as primary when they “define emergent abilities of large language models as abilities that are not present in smaller-scale models but are present in large-scale models; thus, they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models” (Wei et al. “Emergent Abilities,” 2).
- <sup>9</sup> Emily M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610–23.
- <sup>10</sup> Ted Chiang, “ChatGPT Is a Blurry JPEG of the Web,” *New Yorker*, February 9, 2023, [www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web](https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web).
- <sup>11</sup> Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli, “Hallucination Is Inevitable: An Innate Limitation of Large Language Models,” [arXiv:2401.11817](https://arxiv.org/abs/2401.11817).
- <sup>12</sup> Frank Hyneman Knight, *Risk, Uncertainty and Profit* (Houghton Mifflin, 1921).
- <sup>13</sup> James Cameron and Juli Abouchar, “The Precautionary Principle: A Fundamental Principle of Law and Policy for the Protection of the Global Environment,” *Boston College International and Comparative Law Review* 14, no. 1 (1991): 1–27.
- <sup>14</sup> Alex John London, “Threats to the Common Good: Biochemical Weapons and Human Subjects Research,” *Hastings Center Report* 33, no. 5 (2003): 17–25.
- <sup>15</sup> Karthik Valmeekam et al., “On the Planning Abilities of Large Language Models—A Critical Investigation,” paper presented at Advances in Neural Information Processing Systems 36 (NeurIPS 2023).



<sup>16</sup> Eunice Yiu, Eliza Kosoy, and Alison Gopnik, “Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet),” *Perspectives on Psychological Science* 19, no. 5 (2023): 874–83, [doi.org/10.1177/1745691623120140](https://doi.org/10.1177/1745691623120140); Zhijing Jin et al., “Cladder: Assessing Causal Reasoning in Language Models,” in *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

<sup>17</sup> Hyunwoo Kim et al., “FANToM: A Benchmark for Stress-Testing Machine Theory of Mind in Interactions,” [arXiv:2310.15421](https://arxiv.org/abs/2310.15421).

<sup>18</sup> Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, “Are Emergent Abilities of Large Language Models a Mirage?,” paper presented at Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

<sup>19</sup> National Security Commission on Emerging Biotechnology, “White Paper 3: Risks of AixBio,” January 2024, [www.biotech.senate.gov/wp-content/uploads/2024/01/NSCEB\\_AixBio\\_WP3\\_Risks.pdf](https://www.biotech.senate.gov/wp-content/uploads/2024/01/NSCEB_AixBio_WP3_Risks.pdf); Christopher A. Mouton, Caleb Lucas, and Ella Guest, “The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Read-Team Study,” The RAND Corporation, [doi.org/10.7249/RR2977-2](https://doi.org/10.7249/RR2977-2); OpenAI, “Building an Early Warning System for LLM-Aided Biological Threat Creation,” [openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation#results](https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation#results).

<sup>20</sup> Alex John London, *For the Common Good: Philosophical Foundations of Research Ethics* (Oxford University Press, 2023).




1776 EYE STREET, NW, SUITE 1000  
WASHINGTON, DC 20006  
(202) 296-4810  
WWW.NTI.ORG

---

 [FACEBOOK.COM/NTI.ORG](https://www.facebook.com/nti.org)

 [@NTI\\_WMD](https://twitter.com/NTI_WMD)

 [NTI\\_WMD](https://www.instagram.com/nti_wmd)

 [NUCLEAR THREAT INITIATIVE](https://www.linkedin.com/company/nuclear-threat-initiative)

