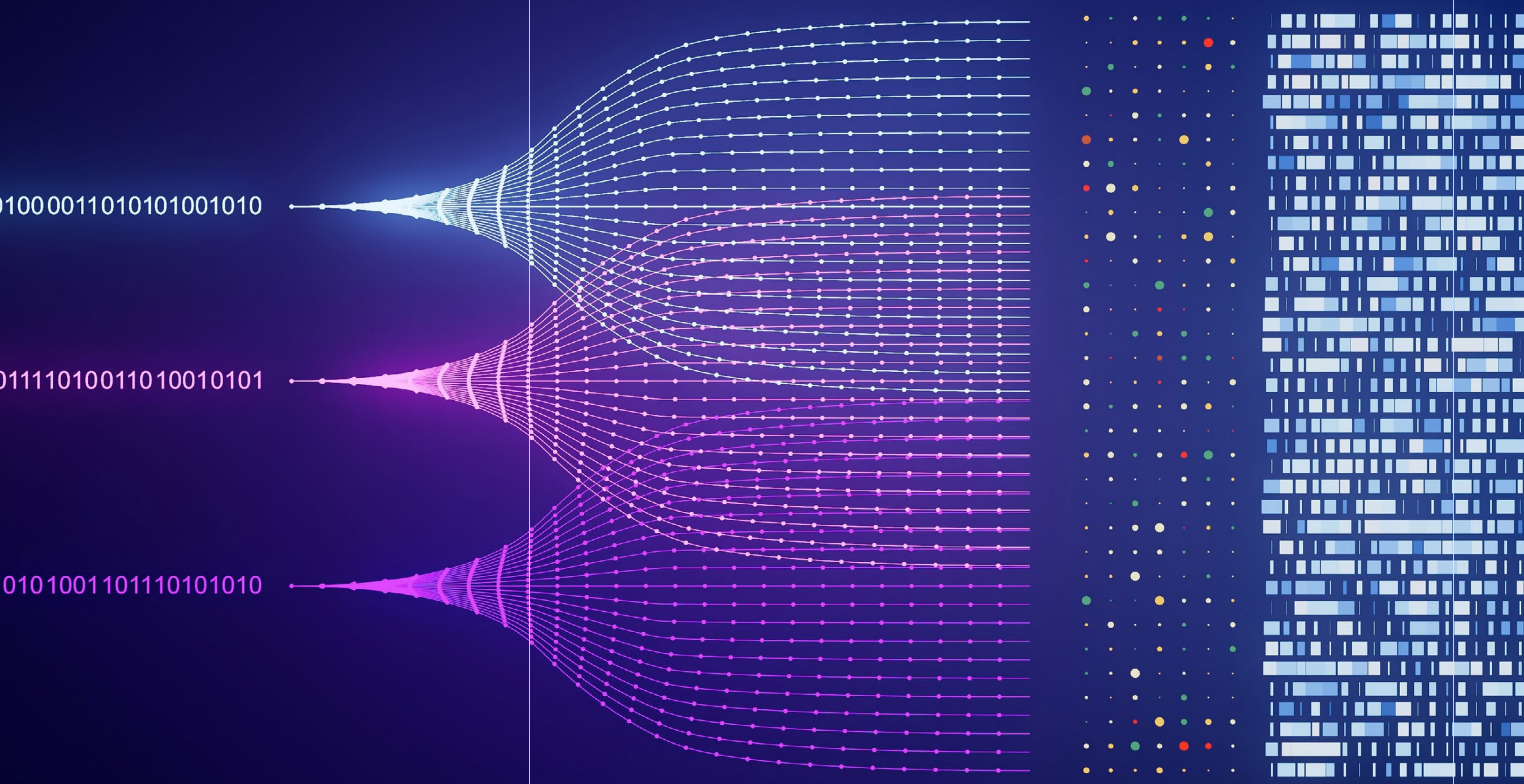
POLICY REPORT
MARCH 2026

Making Machine Learning Safe for the World

BY VINCENT J. CARCHIDI



NEW LINES
INSTITUTE
FOR STRATEGY AND POLICY

NEW LINES INSTITUTE FOR STRATEGY AND POLICY

Our mission is to provoke principled and transformative leadership based on peace and security, global communities, character, stewardship, and development.

Our purpose is to shape U.S. foreign policy based on a deep understanding of regional geopolitics and the value systems of those regions.

CONTENTS

The content and views expressed in this policy report are those of the author and should not be taken to reflect an official policy or position of New Lines Institute for Strategy and Policy.

Cover Image: Illustration depicting AI with interconnected neural net lines and data. (Iurii Krasilnikov / Getty Images)

EXECUTIVE SUMMARY

The productive impacts of state-of-the-art machine learning (ML) have been, and will continue to be, sharply limited given the difficulty or inability to integrate it with safety-critical applications — those in which failures of a system during operation would cause harm to individuals, the public, or the environment. Safety-critical domains are the nexus at which problems that plague state-of-the-art ML in lower-stakes domains meet, revolving principally around matters of reliability, human interpretability, and the ability to intervene in the system's internal mechanisms during operation. That general-purpose ML models today cannot meet the bar for their adoption in safety-critical domains contributes to a sense that they are ever-present with only modest social and economic impacts.

ML, having compelled the world to accommodate it in its most versatile forms, must change if it is to succeed in the most sensitive domains of application. Such changes should follow in the lineage of the most impactful engineering marvels of recent history by providing *guarantees on performance in safety-critical domains*, minimizing the probability of harmful outputs and reducing their severity.

Policy Recommendations

1 INSTRUCT RESEARCH AND DEVELOPMENT

The National Artificial Intelligence Initiative Office (NAIIO), together with the Networking and Information Technology Research and Development (NITRD) Subcommittee, should instruct the National Artificial Intelligence Research & Development Interagency Working Group to coordinate investments in safety-critical neuro-symbolic AI research and development totaling $1.5 billion over a five-year period.

2 INFORM AND HELP GUIDE INVESTMENT

The National Institute of Standards and Technology's (NIST) Center for AI Standards and Innovation (CAISI) should inform — but not exclusively guide — the NAIIO's investment coordination by convening a working group that establishes a new evaluation metrics research agenda specifically for safety-critical neuro-symbolic systems.

3 DETERMINE IF AUTONOMY READINESS LEVELS ARE WARRANTED

A separate working group convened by NIST's CAISI should determine whether Autonomy Readiness Levels (ARLs) for non-defense, safety-critical applications are warranted at the current stage of development. These are evaluation metrics specifically for models expected to perform autonomously over a potentially shifting range of circumstances in the execution of a human-given input. AI "agents" may be considered a part of this target group.

4 NATIONAL SCIENCE FOUNDATION SHOULD ESTABLISH AI RESEARCH INSTITUTE

The National Science Foundation (NSF) should parallel the NAIIO's efforts by establishing a National AI Research Institute dedicated to a specific application area of neuro-symbolic AI. An initial investment worth $80 million to $100 million over five years should be made.

The Safety-Critical Standard

Machine learning (ML) is a success story. This subfield of artificial intelligence (AI) has made short work of the clean-cut distinctions formerly used to make sense of the technology. Gone are the days of "narrow" and "general" models that today appear quaint against the state-of-the-art. This is particularly the case with the advent of generative pre-trained transformers, enabling the proliferation of widely applicable large language models (LLMs) at astonishing speed.

Yet, ML's success is limited, as it is within any scientific or technological paradigm: Complex ML models do not provide guarantees on their performance; they are unreliable. Even the most capable models, when deployed without human supervision in real-world conditions, cannot be trusted to operate safely where their failures would cause harm to either users or the local environment. State-of-the-art ML has thus failed to clear the bar necessary for its trustworthy integration into safety-critical systems.

> "Even the most capable models, when deployed without human supervision in real-world conditions, cannot be trusted to operate safely where their failures would cause harm to either users or the local environment. State-of-the-art ML has thus failed to clear the bar necessary for its trustworthy integration into safety-critical systems."

This outcome is unsurprising. The most relevant trendline in LLM development is the visible improvements along some capability measures — higher benchmark scores, for example — without concomitant improvements in real-world reliability and performance guarantees.[1] This trendline persists.

A National Academies of Sciences, Engineering, and Medicine report forcefully makes the point: So long as ML models cannot clear this bar, the technology in its most capable forms will not live up to the 20th century's greatest engineering accomplishments, including airplanes, automobiles, and nuclear technologies — safety-critical systems whose failures during operation would result in harm, potentially catastrophic, to the user, the public, or the environment.[2]

To achieve this, policymakers should recognize that ML must be made safe *for* the world — a process that may find ML itself lurching into other subfields of AI, undergoing technical changes over time sufficient to designate whatever results as a new paradigm. The burden is technical as much as it is organizational. With its successes limited, ML may not meet the technical standards necessary for safety-critical domains via its own attributes. Thus, resources should be devoted to hybrid AI research, with neuro-symbolic AI representing the most suitable, if broad, candidate.

This report takes no stance on whether AI systems suitable for safety-critical domains are "intelligent" or capable of "reasoning." These are terms that distract rather than clarify, and whose use is entirely unwarranted in making ML systems safe.

Federal government coordination of investments must take up this burden. Bodies like the National Artificial Intelligence Initiative Office should coordinate investments in safety-critical neuro-symbolic research & development totaling $1.5 billion over the next five years. These investments reflect a blunt reality: ML, in its most versatile forms, is not achieving the reliability necessary for its integration into the most sensitive domains of application.

Working groups should likewise be established by the Center for AI Standards and Innovation to inform this investment coordination, focused on a research agenda for safety-critical neuro-symbolic evaluation metrics and the development of Autonomy Readiness Levels for autonomous commercial systems.

Finally, the National Science Foundation should establish a National AI Research Institute devoted to a specific neuro-symbolic application area.

[1] On this, see Carchidi, V. J. (2025) "American AI leadership should not be defined by machine learning." New Lines Institute. https://newlinesinstitute.org/strategic-technology/american-ai-leadership-should-not-be-defined-by-machine-learning/.

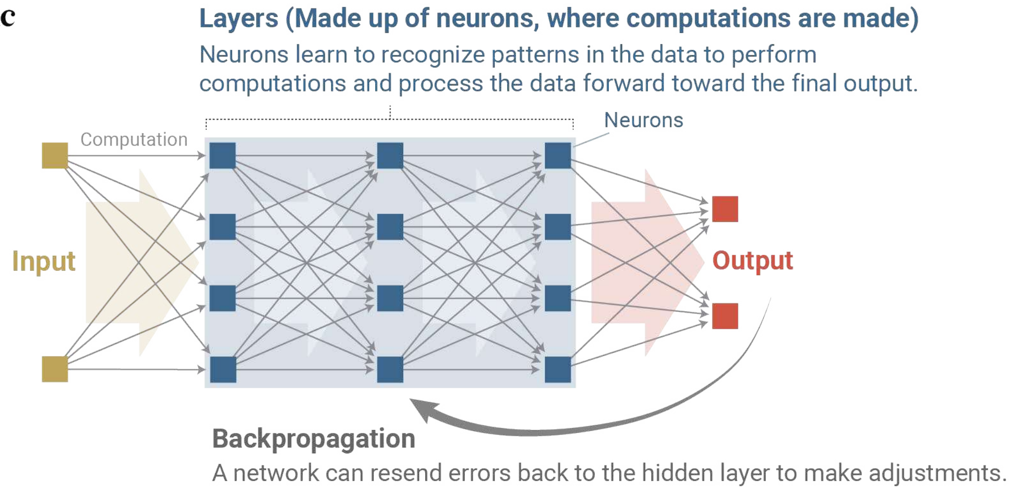
[2] National Academies of Sciences, Engineering, and Medicine. (2025). "Machine Learning for safety-critical applications: Opportunities, challenges, and a research agenda." https://www.nationalacademies.org/publications/27970.

A Crash Course in ML

ML is a subfield of AI. This approach to AI, in its broadest formulation, takes inspiration from biology: The complex networks of interactive neurons in biological (mammalian) brains can be replicated for the end of producing intelligent machine behavior. These neurons can be thought of, simply, as individual units that perform a specific procedure (expressed in mathematical terms). An artificial neural network is thus constructed, at first shallow, consisting of a layer of these artificial neurons at either side, representing "input" and "output." Between them, in the middle, is a single layer of neurons. Having received data from the input layer, a transformation occurs that turns that data into an output. Contemporary neural networks retain this basic design, but are *deep*, made up of *many* layers between input and output.

A Simple, Synthetic Neural Network

The illustration shows a simplified diagram of the complex function of a neural network.


Layers (Made up of neurons, where computations are made)
Neurons learn to recognize patterns in the data to perform computations and process the data forward toward the final output.
Neurons
Computation
Input
Output
Backpropagation
A network can resend errors back to the hidden layer to make adjustments.

Sources: New Lines Institute research including interviews with developers, and a review of books, journals, and technical papers from 2022 and 2023

Deep neural networks are responsible for today's state-of-the-art models across many domains, falling under the umbrella of an ongoing "deep learning revolution."[3]

During training, artificial neural networks *learn* by *statistically associating* data. ML is sometimes described as "data-driven" for this reason. Their learning is guided by an *objective* given to them during training (e.g., to predict the *next word* of an incomplete sentence, to predict the appropriate *designation* for the content of an image, etc.). A neural network learns what a complete sentence "looks like" or what a pine tree "looks like" based on the predictions it makes during training,[4] iteratively updating the connections between artificial neurons based on the accuracy of its prediction to improve itself.

Training techniques vary, though the most relevant for our purposes is the process during which the network predicts *part* of an object (for example, the next word) from its *remainder* (for example, an incomplete sentence where the next word is masked). Consider the following sentence: The dog chased the [masked word]. If the network predicts that the masked word is "tree," this prediction is matched against the actual next word — "car" — and sends a message to its constituent neurons that "The dog chased the" is more likely associated with "car" than "tree." This process repeats over the entire training dataset in what the field refers to as *self-supervised learning*.[5]

[3] For an overview of artificial neural networks' history and their throughline to the present, see, Carchidi, V. J. (2025). "American AI leadership should not be defined by machine learning." New Lines Institute. https://newlinesinstitute.org/strategic-technology/american-ai-leadership-should-not-be-defined-by-machine-learning/, 153-157. For a concise explanation of what constitutes the "deep learning revolution," see generally, Lappin, S. (2025). "The deep learning revolution in AI." European Review, 33(4), 416-425. https://doi.org/10.1017/S1062798725100379.

[4] For a useful breakdown of LLM composition and training specifically, see, Ananthaswamy, A. (2024). "How close is AI to human-level intelligence?" Nature, 636, 22-23. https://doi.org/10.1038/d41586-024-03905-1.

[5] National Academies of Sciences, Engineering, and Medicine. (2025). "Machine Learning for safety-critical applications: Opportunities, challenges, and a research agenda." https://www.nationalacademies.org/publications/27970, 14.

This most basic approach underlies most euphoric investment in AI today: the generative pre-trained transformer, or GPT. The GPT is responsible for the LLMs of great commercial interest. LLMs are thus part of the broader ML family, currently representing a class of state-of-the-art models.

Performance (Not) Guaranteed

An LLM *can* produce coherent text in natural language, writing essays on any subject to which you set it, but it will sometimes produce plausible but incorrect, misleading, or fictitious outputs (often of an inhuman kind; "hallucinations"). A multimodal model *can* produce images or video of startlingly realistic quality, but the physical relations between objects and people will invariably lack grounding in reality. Self-driving vehicles *can* autonomously maneuver within unstructured environments, but the more unfamiliar and exotic the environment, the more likely they will be to crash in inhuman ways or freeze in the middle of the road.[6]

> "All these are examples of complex ML models or the systems they underpin failing to provide performance guarantees; that is, reliability of performance sufficient to deploy in real-world conditions without regular human intervention or oversight and consistent with the intended applications of the model."

All these are examples of complex ML models or the systems they underpin failing to provide *performance guarantees*; that is, reliability of performance sufficient to deploy in real-world conditions without regular human intervention or oversight and consistent with the intended applications of the model. Real-world reliability in ML models is a long-recognized problem.[7] To be clear, there are two sides to guaranteed performance: *accuracy* and *consistency*.[8] To suggest that a model clears the bar for integration with a safety-critical system is to suggest, minimally, that a model's outputs are accurate with respect to its intended application and that the same input will yield the same output across multiple instances (this is sometimes called *determinacy*).

Defense analyst Mary Cummings argues that generative AI systems like LLMs fall short on both counts: They are unacceptably inaccurate (providing misleading or faulty outputs) and unacceptably inconsistent (the same input may not yield the same output).[9] Specifically, Cummings argues that generative AI systems do not engage in a form of higher reasoning sufficient to deal with conditions of greater uncertainty — a fact about generative AI systems that, she argues, makes them unsuitable for *any* safety-critical system. (Cummings' central argument is that generative AI should not be permitted to influence the control, activation, supervision, direction, or governance of any *weapon* system.)

Perhaps most useful for our purposes is Cummings' observation about the development and deployment of generative AI models. In contrast to drone technology — which underwent sustained military research and development for 30 years, generative AI underwent zero such years. Its development and subsequent rapid commercialization, she argues, came at the expense of learning from its early failures to better understand its limitations.[10] Cummings here invokes the U.S. Department of Defense's (DoD) Technology Readiness Levels (TRLs):[11]

[6] A power outage in December 2025 in San Francisco — among the limited number of cities where Waymo self-driving vehicles are licensed for operation — illustrated the point: a number of Waymo vehicles froze in the streets with their hazard lights on when confronted with non-operational traffic signals, prompting the company to halt operations for that day. The company's "human fleet" of remote operators appear to have been overwhelmed by the abruptness and widespread nature of the failures. See, Roy, A. (2025, December 27). "Waymo's San Francisco outage raises doubts over robotaxi readiness during crises." Reuters. https://www.reuters.com/business/autos-transportation/waymos-san-francisco-outage-raises-doubts-over-robotaxi-readiness-during-crises-2025-12-27/.

[7] See, Grote, T., Genin, K., & Sullivan, E. (2024). "Reliability in machine learning." Philosophy Compass, 19(5), 1-11. https://doi.org/10.1111/phc3.12974.

[8] Carchidi, V. (2025, September 22). "What is artificial intelligence? A primer." Defense and Security Monitor. https://dsm.forecastinternational.com/2025/09/22/what-is-artificial-intelligence-a-primer/.

[9] See generally, Cummings, M. (2025). "Prohibiting generative AI in any form of weapon control." NeurIPS 2025 Position Paper Track, 1-13. https://openreview.net/forum?id=uEY7kQsiZz&referrer=%5Bthe%20profile%20of%20Mary%20Cummings%5D(%2Fprofile%3Fid%3D~Mary_Cummings1).

[10] Cummings, M. (2025). "Prohibiting generative AI in any form of weapon control." NeurIPS 2025 Position Paper Track, 8-10. https://openreview.net/forum?id=uEY7kQsiZz&referrer=%5Bthe%20profile%20of%20Mary%20Cummings%5D(%2Fprofile%3Fid%3D~Mary_Cummings1).

[11] For a breakdown, see the following: https://api.army.mil/e2/c/downloads/404585.pdf.

A humanoid robot practices grasping objects at the data collection training ground for humanoid robots in Qingdao, Shandong, China, on Jan. 12, 2026. (Costfoto/NurPhoto via Getty Images)

Levels 1-3: Academic Research (basic conceptual research and proof of concept)

Levels 4-6: Technological Maturation (testing and validation in low- and higher-fidelity environments)

Levels 7-9: System Integration (prototype demonstration in an operational environment, validation, and ultimate application)

It is at the middle TRLs — Technological Maturation — during which generative AI's development was short-circuited, Cummings argues.

As the TRLs indicate, threats to model reliability include more than the level of model development in laboratory settings. They also include the *deployment* of the model and the *organizational structure* that accommodates the model. As Grote, Genin, and Sullivan detail:

1. **Model Development:** There is no established theory of how deep learning models generalize;
2. **Model Deployment:** Models are often used in unstable environments to which they are vulnerable;
3. **Adaptation of the Organization's Environment:** The model's alignment with the environment it is in depends on other data beyond what it is trained on and is hampered by opaque operations.[12]

Arguably, much of the burden of making ML models suitable for real-world deployments since roughly 2023 has been placed on the third factor: the organization is to find creative ways of accommodating the model, warts and all. Workflows are restructured around poorly understood limitations. However, as this occurs, the first two factors are comparatively lessened in meeting this burden, failing to establish adequate processes for verifying and validating the model's performance within intended and unintended conditions *before* deployment — harkening back to the short-circuiting of generative AI's technological maturation.

In any event, the authors note that what ties together issues related to the

[12] Grote, T., Genin, K., & Sullivan, E. (2024). "Reliability in machine learning." Philosophy Compass, 19(5), 1-2. https://doi.org/10.1111/phc3.12974.

robustness of models is the *optimal aggregation* of human and machine judgment.

Although these components relevant to reliability were recognized and developed prior to the rise of general-purpose models, LLMs have not changed the calculus. Indeed, they have, if anything, *compounded* these threats to reliability: the jump in capabilities, rather than tested and validated in a period of technological maturation, instead led to abrupt and widespread deployment.

Counterbalancing this, however, is that LLMs have tended to be deployed in lower-stakes domains where the risk of failure is not exposed to critical functions. As Narayanan and Kapoor observe, the diffusion of innovative technologies tends to lag in safety-critical domains, with today's LLMs largely absent, owing to the difficulty of preempting the diversity of possible deployment conditions for such complex models during testing and validation.[13] The risks responsible for this lag are noticed in a recent report on AI safety management. Authors Eric Marsden and Véronique Steyer note:

> *While there are numerous expected benefits, such as the ability to process large volumes of reliability data or unstructured natural language incident reports, the structural opacity of large neural networks, their nondeterministic nature, and their capacity to learn from new data mean that traditional safety assurance techniques used for conventional software are not applicable*.[14]

They similarly observe that human-machine collaboration with contemporary AI agents "introduce new variables" in safety considerations.[15]

Indeed, new risks carry a need for new analyses of ML models: how they are "designed, evaluated, and ultimately deployed in real-world settings," Kamal Acharya and Houbing Song argue, require a "critical reassessment."[16] They usefully analogize a trustworthy model in safety-critical domains to the design of a car for *all terrains*. This vehicle can perform consistently on *smooth* roads (clean, anticipated data), *bumpy* roads (irrelevant or faulty data), within *inclement weather* (unexpected, unfamiliar data), and amid *GPS tampering* (adversarial attacks, or deliberate attempts at sabotage). Building such a system impacts both technical design and testing and validation.

Reliability remains a matter of optimally aggregating human and machine judgment. However, after having the world scramble to accommodate LLMs, the burden of such aggregation nevertheless has disproportionately fallen on organizations, accounting in part for their apparent widespread presence alongside their relatively modest social and economic impacts. The burden should therefore be tilted back toward research and development that takes seriously the design of complex ML models, the theories employed to account for their behavior, and testing and validation therein.

Classical AI's Moment in the Sun

The reliance that ML models have on statistical associations of data is a corollary for their difficulty integrating into safety-critical domains, where the "safety gap between 95 and 99.99999999 percent is very large and essentially impossible to bridge through statistical learning alone,"[17] as the National Academies report bluntly states.

[13] Narayanan, A. & Kapoor, S. (2025, April 15). "AI as normal technology." Knight First Amendment Institute. https://knightcolumbia.org/content/ai-as-normal-technology, 4.

[14] Marsden, E. & Steyer, V. (2025). "AI and safety management: An overview of key challenges." https://www.foncsi.org/sites/default/files/Publications/Publications_EN/CSI_Foncsi_202503_AI-safety-management-overview.pdf, v.

[15] Ibid.

[16] Acharya, K. & Song, H. (2025). "A comprehensive review of neuro-symbolic AI for robustness, uncertainty quantification, and intervenability." Arabian Journal for Science and Engineering, 1. https://doi.org/10.1007/s13369-025-10887-3.

[17] National Academies of Sciences, Engineering, and Medicine. (2025). "Machine Learning for safety-critical applications: Opportunities, challenges, and a research agenda." https://www.nationalacademies.org/publications/27970, 24.

To be sure, AI contains diverse approaches. The euphoria that defines investments in ML today was once reserved for another class of techniques: symbolic AI, sometimes called Good Old-Fashioned AI. Revisiting it will sober our views on the reliability issues present in contemporary ML and inform our views of what is technically necessary going forward.

If ML is premised on building machines that learn through data, then symbolic AI can be usefully distinguished as systems *hand-coded with human knowledge*; their knowledge is structured by humans. This knowledge is represented via *symbols*. These symbols can be words, rules, or logical statements. The system manipulates symbols to determine relationships between them via the knowledge it has been given by humans.[18]

Symbolic systems are not scalable in any manner comparable to their ML counterparts. They are thus *highly restricted* in application; exceed the scope of their human-given knowledge and their performance collapses. Nevertheless, symbolic systems are notable for their ability to provide *performance guarantees* within those contexts. They offer, in narrowly tailored domains, accuracy and consistency of output. Humans can also audit the outputs of a symbolic system[19] to make sense of them within context (interpret them) and understand how those outputs resulted from the input (explain them).

So far as safety-critical domains are concerned, the primary difference between ML and symbolic systems is in the *reliability-flexibility tradeoff*: LLMs in particular can effectively deal with diverse datasets at scale but without provable correctness and predictable reliability. Symbolic systems, in contrast, guarantee accuracy and consistency but only within narrow and well-defined domains of application.[20] The limits of both forms of AI's societal impacts are defined by the extent to which their outputs can be uncritically relied upon and investigated or intervened upon if necessary.

This tradeoff is an old problem. During the Defense Advanced Research Projects Agency's (DARPA) 10-year, next-generation computing effort — the Strategic Computing Initiative of 1983-1993 — symbolic-based systems were one of the central means by which the agency sought to adjust to the quickening tempo of modern battles (emphasizing "smart" computerized weapon systems and forces). DARPA envisioned the construction and deployment of service-specific, autonomous, and intelligent systems.[21]

However, as some critics wrote in the Bulletin of Atomic Scientists in 1984, reliance on computerized systems in mission-critical domains "creates a false sense of security" as AI systems' propensity to "act inappropriately in unanticipated situations" owes to a "fundamental limit on their reliability."[22] The mismatch between system design and humans' inability to "anticipate all the circumstances they will encounter" results, they argue, in "a much more intractable kind of failure…"[23]

[18] For a brief overview, see, Carchidi, V. J. (2025). "American AI leadership should not be defined by machine learning." New Lines Institute. https://newlinesinstitute.org/strategic-technology/american-ai-leadership-should-not-be-defined-by-machine-learning/, 151-153.

[19] Feldstein, J., Dilkas, P., Belle, V., & Tsamoura, E. (2024). « Mapping the Neuro-Symbolic AI landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning." ArXiv, 3. https://doi.org/10.48550/arXiv.2410.22077.

[20] Jovanović, M. & Campbell, M. (2025). "Reasoning AI: Progress, challenges, and the path forward for large reasoning and concept models." Computer, 58(11), 118. https://doi.org/10.1109/MC.2025.3596671.

[21] See generally, Defense Advanced Research Projects Agency. (1983). "Strategic computing: New-Generation computing technology: A strategic plan for its development and application to critical problems in defense (NTIS Issue No.198419)." National Technical Reports Library, U.S. Department of Commerce. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/ADA141982.xhtml.

[22] Ornstein, Severo M., Smith, Brian C., & Suchman, Lucy A. (1984). "Strategic computing." Bulletin of the Atomic Scientists, 40(1), 12. https://doi.org/10.1080/00963402.1984.11459292.

[23] Ibid., 13.

People visit the NVIDIA booth at an exhibition in Hangzhou, Zhejiang Province, China, in 2023. (Costfoto/NurPhoto via Getty Images)

Recent Trends in ML (Un)reliability

The general observation by Strategic Computing's critics is not new. In his 2018 critical appraisal of deep learning, Gary Marcus explicitly noted that

> *"it is comparatively easy to make systems that work in some limited set of circumstances (short term gain), but quite difficult to guarantee that they will work in alternative circumstances with novel data that may not resemble previous training data (long term debt, particularly if one system is used as an element in another larger system)."*[24]

This emphasis on reliability in novel circumstances — and ML models' persistent failures therein — is remarkably consistent.[25]

In October 2022, one month before the release of OpenAI's ChatGPT-3.5, an article written by Don Monroe appeared in the Communications of the ACM noting that "deep learning does not provide the sorts of performance guarantees that are customary in computer science."[26] Neuro-symbolic AI, an approach that blends both ML and symbolic techniques, was a noted contender for this problem's mitigation.

In mid-November 2022, two weeks before ChatGPT-3.5's release, Meta released "Galactica," described by researchers as a "large language model for science."[27] Galactica was an early attempt to develop systems capable of accurately summarizing scientific texts in natural language to aid human scientists. The model was quickly pulled, having output a litany of then-bizarre hallucinations, a phenomenon with which the public was not yet acquainted, as Sharon Goldman retrospectively details.[28]

Two weeks later, the release of ChatGPT-3.5 triggered a corporate competition over this form of conversational AI, with the web-based app garnering an estimated 100 million active monthly users by January 2023.[29]

Interestingly, the standards to which LLMs were held among industry commentators became shockingly tolerant of errors and unreliability, often with promises that the solution to hallucinations is within reach.

Then co-founder of Inflection AI and current CEO of Microsoft AI Mustafa Suleyman posted on Twitter in January 2023: "LLM hallucinations will be largely eliminated by 2025."[30] Having apparently changed his view, Suleyman in April 2025 told an interviewer that hallucinations are "solutions" to the problem of traditional database storage in that the stored data undergoes a novel transformation (i.e., a source of creativity).[31] Alphabet CEO Sundar Pichai, for his part, noted in September 2023 that "[i]n the near-term, state-of-the-art LLMs have hallucination problems — they can make up things."[32] In November 2025, his position largely unchanged, he cautioned against "blind trust" in generative AI output.[33]

Equally interesting, consumer usage of the apps hosting these models increased dramatically over the same period. OpenAI noted in September 2025

[24] Marcus, G. (2018). "Deep learning: A critical appraisal." ArXiv, 14. https://doi.org/10.48550/arXiv.1801.00631.

[25] Marcus, G. (2025, June 15). "AI's reliability crisis." Project Syndicate. https://www.project-syndicate.org/magazine/generative-ai-fundamentally-unreliable-and-with-no-apparent-solution-by-gary-marcus-2025-06.

[26] Monroe, D. (2022, October 1). "Neurosymbolic AI." Communications of the ACM. https://cacm.acm.org/news/neurosymbolic-ai/.

[27] Taylor, R. et al. (2022). "Galactica: A large language model for science." ArXiv, 1-58. https://arxiv.org/abs/2211.09085.

[28] Goldman, S. (2023, November 14). "What Meta learned from Galactica, the doomed model launched two weeks before ChatGPT." VentureBeat. https://venturebeat.com/ai/what-meta-learned-from-galactica-the-doomed-model-launched-two-weeks-before-chatgpt.

[29] Hu, K. (2023, February 1). "ChatGPT sets record for fastest-growing user base — analyst note." Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[30] See here: https://x.com/mustafasuleyman/status/1667184880235446280?lang=en.

[31] See here: https://www.youtube.com/watch?v=cOzbBIV-frs.

[32] Levy, S. (2923, September 11). "Sundar Pichai on Google's AI, Microsoft's AI, OpenAI, and…did we mention AI?" Wired. https://www.wired.com/story/sundar-pichai-google-ai-microsoft-openai/.

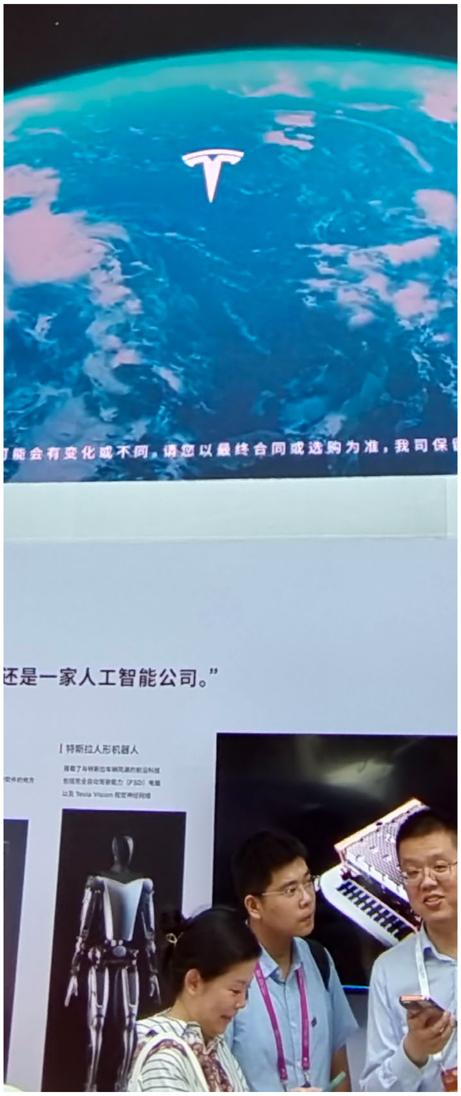
[33] Islam, F., Clun, R., & McMahon, L. (2025, November 18). « Don't blindly trust what AI tells you, says Google's Sundar Pichai." BBC. https://www.bbc.com/news/articles/c8drzv37z4jo.

that ChatGPT had 700 million weekly active users.[34] Google noted that the Gemini app had 650 million monthly active users in November 2025.[35] This rapid proliferation and, at least, individual usage of consumer-facing LLMs has thus occurred despite the models' lack of "customary"[36] performance guarantees. Unreliability has not deterred usage, albeit largely of a non-safety-critical kind.

Currently, over three years into the generative AI boom, models routinely fail for their intended purposes. Euphoria over general-purpose models has neglected the fact that — charitably speaking — "general-purpose" might once have meant that an LLM can faithfully execute any task to which the model is set at an expected level of performance within a wide-range of anticipated and unanticipated set of circumstances. Yet GPT-based systems do not offer this. You can train and set them to any task for which there is sufficient data, but they will not faithfully execute them at a consistent level of performance. They will and do fail in arbitrary circumstances.

Note that these failures do *not* mean state-of-the-art ML models are useless. Rather, our concern with safety-critical applications merely brings this lack of performance guarantees to the fore, as it is within these domains — in which individuals, the public, or the environment could be harmed, potentially catastrophically — where failures during operation are the least tolerable. Yet the deficiencies that could lead to such failures are pervasive across state-of-the-art ML models.

All this is rather unlike societally transformative technologies taken completely for granted**.** When you start the engine of your car, a series of controlled explosions are unfolding under the hood — largely out of view, out of mind, save for some occasional car trouble. Most cars do not blow up unexpectedly. They offer performance guarantees, in more ways than one. The same holds for airplanes, boats, complex medical devices, drug manufacturing, and any other system whose performance is both safety-critical and hinges on a multitude of subsystems operating in tandem. When these systems *do* fail catastrophically, it is typically within a remarkably small margin of error and sometimes because of perverse regulatory or market incentives that distort design, testing, and validation processes.


可能会有变化或不同，请您以最终合同或选购为准，我司保
还是一家人工智能公司。"
特斯拉人形机器人
搭载了与特斯拉车辆同源的前沿科技
包括完全自动驾驶能力（FSD）电脑
以及 Tesla Vision 视觉神经网络

Visitors visit a Tesla booth at the 2023 World Artificial Intelligence Conference in Shanghai, China. (Costfoto / NurPhoto via Getty Images)

Three Sources of LLM Unreliability

Crucially, the capabilities of LLMs have *improved along certain measures* during the generative AI boom, the specifics of which are sometimes shocking. The clearest view of LLM progress is provided by ever-increasing benchmark scores, often necessitating the creation of new benchmarks. However, these improvements have not resulted in levels of reliability sufficient for deployment and unsupervised operation in safety-critical domains. Nor do benchmark scores transfer seamlessly to real-world applications. Most importantly, there is no indication that the research tracks LLMs are currently on are trending in the direction necessary to allay fears of their unreliability.

This is a significant claim. Thus, to support this claim, three distinct, if overlapping, sources of evidence are brought to bear:

1. **Improvements Without Guarantees**: Improvements in capabilities do not typically lead to performance guarantees on benchmark testing.

[34] OpenAI. (2025, September 15). "How people are using ChatGPT." https://openai.com/index/how-people-are-using-chatgpt/.

[35] Elias, J. (2025, November 18). "Google announces Gemini 3 as battle with OpenAI intensifies." CNBC. https://www.cnbc.com/2025/11/18/google-announces-gemini-3-as-battle-with-openai-intensifies.html.

[36] Monroe, D. (2022, October 1). "Neurosymbolic AI." Communications of the ACM. https://cacm.acm.org/news/neurosymbolic-ai/.

2. **Benchmark Score Instability**: Improvements in benchmark scores are themselves unstable, depending in part on the variation of the test and its design, merely adding to uncertainty about operations during model deployment.
3. **Difficulty Transferring to Reality**: Benchmark scores, even where stable, do not seamlessly or predictably transfer to real-world domains.

Taken together, each source of evidence implicates state-of-the-art ML models in their *current and foreseeable* unsuitability for safety-critical applications.

Benchmark Improvements Without Guarantees

Consider research on LLMs' abilities to formulate and execute plans, where the disconnect between performance *improvements* and performance *guarantees* shows clearly.

A 2022 study by an Arizona State University (ASU) research group found that GPT-3 exhibited "dismal performance"[37] on benchmarks testing planning and reasoning capabilities. In 2024, a follow-up test was conducted on successor model GPT-4 on the "Blocksworld" benchmark — a test that requires the model generate plans for stacking blocks. GPT-4 scored approximately 35% accuracy (averaging a 12% success rate in producing executable plans across domains).[38] Performance was, overall, unimpressive.

The introduction of OpenAI's "o-series" models, beginning with o1-preview in September 2024,[39] was a watershed. The same ASU research group tested o1-preview on three variants of the Blocksworld test. On a version with complete knowledge of the problems, it scored a remarkable 97.8% accuracy. On a version with incomplete knowledge, its score dropped to a still-impressive 52.8% accuracy. Its poorest result of 37.3% was on an altered and randomized version of the test.[40] At the time, these results far and away overshadowed other LLMs, like Claude 3.5 Sonnet.

"At least on benchmark scores ... LLMs improved between 2022 and 2024. However, this improvement, though directionally significant, provides little confidence in their suitability for safety-critical domains, even with the expectation of future improvements."

At least on benchmark scores such as these, LLMs improved between 2022 and 2024. However, this improvement, though directionally significant, provides little confidence in their suitability for safety-critical domains, even with the expectation of future improvements.

The same study came with a twist: The ASU group also tested a symbolic planning system called Fast Downward.[41] This system achieved *100% accuracy on all Blocksworld planning problems*, including those with complete and incomplete knowledge as well as alterations and randomizations. As the researchers note, classical planners like Fast Downward manage this in "a fraction of the time, compute, and cost, while providing guarantees that their answers are correct."[42]

A follow-up study in October 2024 by the ASU group increased the *sizes* of the planning problems (i.e., increased the optimal number of steps to the problems' solutions).[43] OpenAI's o1-preview dropped from an original score of 97.8% on the unaltered Blocksworld test to only 23.63% on problems where at least 20 steps are needed. In sharp contrast, Fast Downward *predictably*

[37] Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). "Large Language Models still can't plan (A benchmark for LLMs on planning and reasoning about change)." ArXiv, 2. https://arxiv.org/abs/2206.10498v1.

[38] Valmeekam, K., Marquez, M., Sreedharan, S., & Kambhampati, S. (2023). "On the planning abilities of Large Language Models: A critical investigation." ArXiv, 5. https://arxiv.org/abs/2305.15771v2.

[39] See OpenAI's announcement here: https://openai.com/o1/.

[40] Valmeekam, K., Stechly, K., Kambhampati, S. (2024). "LLMs still can't plan; can LRMs? A preliminary evaluation of OpenAI's o1 on PlanBench. ArXiv," 1-3. https://doi.org/10.48550/arXiv.2409.13373.

[41] See, Helmert, M. (2011). "The Fast Downward planning system." ArXiv, 1-56. https://doi.org/10.48550/arXiv.1109.6051.

[42] Valmeekam, K., Stechly, K., Kambhampati, S. (2024). "LLMs still can't plan; can LRMs? A preliminary evaluation of OpenAI's o1 on PlanBench. ArXiv," 6. https://doi.org/10.48550/arXiv.2409.13373.

[43] Valmeekam, K., Stechly, K., Gundawar, A., & Kambhampati, S. (2024). "Planning in strawberry fields: Evaluating and improving the planning and scheduling capabilities of LRM o1." ArXiv, 8-10, https://doi.org/10.48550/arXiv.2410.02162.

scaled to harder problems within the planning domain, again achieving 100% accuracy on all problems.

The picture that emerges is that LLMs have improved along specific measures, but these improvements typically do not lead to performance guarantees *within* a given benchmark, nor do these improvements *reliably* or *predictably scale* to harder problems of the same class. State-of-the-art ML models therefore retain, rather than overthrow, the earlier distinction between ML and symbolic AI: Where neural networks can scale to different types of problems and improve their performance over time, they lack the domain-specific reliability of their symbolic counterparts.

For safety-critical deployments, any model purportedly set to one class of tasks must complete these tasks at a rate of accuracy that allows it to operate without constant human supervision. It must likewise be robust to changes in the immediate environment, such that its performance does not degrade catastrophically. LLMs do not make this promise.

Benchmark Score Instability

Benchmarks in ML go through a now-familiar cycle: First, the benchmark is introduced. After some time, at least one model begins to reach 90% or higher accuracy. At this point, upon reaching near-ceiling accuracy, the benchmark's usefulness is diminished as a test for frontier capabilities. (The field calls this "saturation.")As a result, a new benchmark — often with more complex problems of the same domain (e.g., more advanced mathematics problems) — is introduced.

This rinse-and-repeat cycle means that model *reliability* is not a standard capability metric; how the model performs in response to the same input or variations of the input is not adequately assessed.

One February 2025 study targeted model reliability, specifically in the domain of safety-critical applications. Researchers asked: On what kinds of tasks are frontier LLMs *actually* reliable?[44] They studied this by *reexamining* LLMs' performances on older, since-conquered benchmarks. The researchers introduced "platinum"[45] benchmarks in which a 100% score is possible during re-examination.[46]

For most benchmarks, none of the models achieved 100% accuracy. Notably, the researchers found that almost every model made *simple* mistakes on almost every platinum benchmark,[47] except for quite simple test datasets (e.g., SingleOP and MultiArith). For models tested that were more reliable (and more capable overall), two failure mode patterns nonetheless emerged that impacted reliability:

1. **First Event Bias:** Gemini 1.5 Flash, Gemini 1.5 Pro, and Mistral Small fail on over 85% of problems, primarily selecting the first event in a question phrased according to the form: "What happened second: [some event] or [some other event]?" Interestingly, the model's reasoning would sometimes acknowledge the proper order of events, only to answer with the (incorrect) first event.
2. **Rounding Up Primes:** 20% of the time, Claude 3.5 Sonnett rounded up a number even if it was already a whole number, particularly if it was closer to a prime number.[48]

[44] Vendrow, J., Vendrow, E., Beery, S., & Madry, A. (2025). "Do large language model benchmarks test reliability?" ArXiv, 1, https://arxiv.org/abs/2502.03461.

[45] Ibid., 1-2.

[46] They do this in part, though not only, by removing the *ambiguity* previously in certain questions.

[47] Vendrow, J., Vendrow, E., Beery, S., & Madry, A. (2025). "Do large language model benchmarks test reliability?" ArXiv, 5-7, https://arxiv.org/abs/2502.03461.

[48] Vendrow, J., Vendrow, E., Beery, S., & Madry, A. (2025). "Do large language model benchmarks test reliability?" ArXiv, 7-8, https://arxiv.org/abs/2502.03461.

Note the *subtlety* of these mistakes and how they might be uncritically perceived by a human operating in a safety-critical domain, particularly the mismatch between the model's reasoning and its answer in the first example.

In a similar vein, an April 2025 study targeted the *reproducibility* of benchmark scores, focusing specifically on mathematical reasoning benchmarks (the field's golden geese, as it were).[49] The study found "a surprising degree of sensitivity… to seemingly minor design choices" made when testing LLMs.[50] Nine models, grouped into two size categories,[51] were tested on three mathematics benchmarks: AIME'24, AMC'23, and MATH500.[52]

Interestingly, "non-obvious factors" including the hardware underlying the model and the evaluation framework used were among the sources of score instability.[53] For instance, testing the same model across five different graphics processing unit (GPU) clusters on AIME'24 led to performance variations of up to 8% for the tested models, including a version of DeepSeek.[54]

Models likewise show varying sensitivity to setting a maximum word (token) output. The format of the prompt used to pose math problems was an additional source of instability.[55] Whereas some models "rely heavily on format alignment,"[56] others, like Qwen-2.5-Math, benefitted from a prompt-*free* setup (i.e., the question is posed without any accompanying template).[57] Sensitivities leading to score instability were therefore both unusual and varied across models.

This focus is echoed elsewhere. In particular, a March 2025 study showed that newer models performed increasingly well on mathematics benchmarks that scoreed *only* the final numerical answer.[58] When tested on both the final answer and the *proofs* generated to support those answers on 2025 USA Math Olympiad problems, the models exhibited "critical failure modes, including flawed logic, unjustified assumptions, and a lack of creativity."[59]

Building on this, a June 2025 study introduced a new mathematics test dataset consisting of 200 problems specifically targeting models like Google's Gemini-2.5 and OpenAI's o3.[60] The study found that models "frequently fall into single-step reasoning errors, exhibit vague argumentation, utilize non-existent premises, and produce incomplete proofs."[61] It judged that the models' inability to self-reflect produces "subtle logical inconsistencies."[62]

The reader may at this point experience whiplash. A variant of Gemini Deep Think did, after all, win gold-medal standard at the International Mathematics Olympiad in July 2025 (a pre-university level competition).[63] However, this is consistent with concerns about safety-critical reliability: This competition was won by having provided correct answers to only six (difficult) mathematics

[49] Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., & Bethge, M. (2025). "A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility." ArXiv, 1-39, https://arxiv.org/abs/2504.07086.

[50] Ibid., 2.

[51] These were 1.5 billion and 7 billion parameters models.

[52] These datasets can be accessed via the hyperlinks: AIME'24, AMC'23, and MATH500.

[53] Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., & Bethge, M. (2025). "A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility." ArXiv, 7, https://arxiv.org/abs/2504.07086.

[54] These models were OpenRS-1.5B and DeepSeek-R1-Distill-7B.

[55] Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., & Bethge, M. (2025). "A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility." ArXiv, 7, https://arxiv.org/abs/2504.07086.

[56] Ibid.

[57] Ibid., 8-9.

[58] Petrov, I. et al. (2025). "Proof or bluff? Evaluating LLMs on 2025 USA Math Olympiad." ArXiv, 1-13, https://arxiv.org/abs/2503.21934v1.

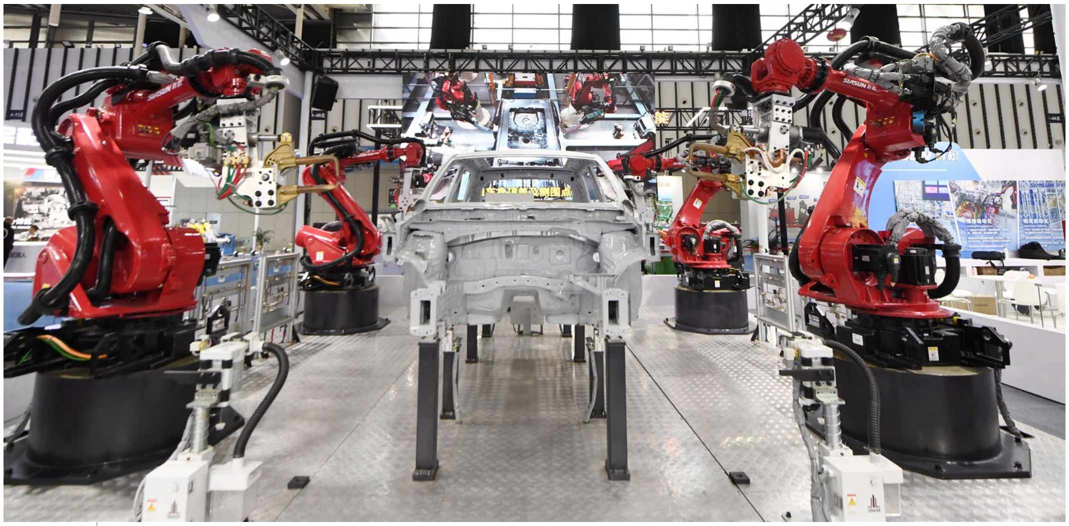
[59] Ibid., 1.

[60] Guo, D., Liu, J., Fan, Z., He, Z., Li, H., Wang, Y., R., Yi, & Fung, M. (2025). "Mathematical proof as litmus test: Revealing failure modes of advanced large reasoning models." ArXiv, 1-41, https://arxiv.org/abs/2506.17114.

[61] Ibid., 8.

[62] Ibid., 3.

[63] Luong, T. & Lockhart, E. (2025, July 21). "Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad." Google DeepMind. https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/.



Industrial robots are on display in an exhibition area at the 2023 World Intelligent Manufacturing Conference in Nanjing, China. (Costfoto / NurPhoto via Getty Images)

problems in a relatively closed domain, but deployment conditions are more complex than highly restricted tests such as this. We may view Gemini Deep Think's score here as the saturation of a particular benchmark, but as we have seen, this does not indicate reliability.

Still, evidence for benchmark instability is continuously arising. An October 2025 study sought to understand how the use of specific characters in prompts impacts the evaluation's results.[64] During benchmark evaluations, prompts are separated. A *single character* (e.g., a comma, a new line, a semi-colon, a hashtag, etc.) can be used for this purpose. This sets the boundaries between problem examples.[65]

The researchers employ a unified evaluation framework and use a single character to separate demonstration examples without changing the *content* of existing benchmark problems. The models tested primarily include open-source language models of varying sizes[66] from the Llama, Gemma, and Qwen families. They were tested on three main benchmarks: MMLU, ARC-Challenge, and Commonsense-QA.[67]

Findings indicate significant instability. Across the three benchmarks, the smaller models' performance fluctuated significantly depending on the character used between examples, with this sensitivity occurring down to the level of individual topics within MMLU, such that *modification of the character could radically shift model rankings*.[68] Put one way: The accuracy of any model's outputs in any arbitrary evaluation instance is essentially a snapshot of a moving target; a distortion of capability.

Importantly, scaling these models from their smaller to larger sizes improves performance on all three benchmarks — *but not robustness*. Indeed, the larger Llama-3.1-70B is *more brittle* than its smaller, Llama-3.1-8B counterpart in the face of character modification. This includes a 40% drop in accuracy from the use of "!" to separate examples to the use of "?" on Commonsense-QA. This finding is replicated with closed-source GPT-4o on MMLU, with a 45.6% difference in accuracy between its highest and lowest scores across character modifications.[69]

[64] Su, J., Zhang, J., Ullrich, K., Bottou, L., & Ibrahim, M. (2025). "A single character can make or break your LLM evals." ArXiv, 1-22, https://arxiv.org/abs/2510.05152.

[65] Ibid., 1.

[66] The model sizes ranged between 8 billion parameters and 70 billion parameters.

[67] Su, J., Zhang, J., Ullrich, K., Bottou, L., & Ibrahim, M. (2025). "A single character can make or break your LLM evals." ArXiv, 1, https://arxiv.org/abs/2510.05152.

[68] Ibid., 2.

[69] Ibid., 6.

Difficulty Transferring to Reality

Benchmark score instability reveals a tendency in AI evaluations: Once a model has conquered a benchmark, the capability required to achieve this is undisputed. Having been earned, the field can move on to bigger and better capabilities. Yet scores are unstable. This prefigures another problem: Model capabilities exhibited on benchmarks do not seamlessly or predictably transfer to real-world deployments.

Some researchers situate their work in the context of *reliability as a deployment metric* and seek to contribute to "reliability quantification" for LLMs.[70] Quantification is no small matter: The transfer of benchmark performance — even where stable — to real-world situations is not guaranteed. Roboticist Rodney Brooks correctly noted that witnessing an AI system performing a task leads humans to "immediately estimate its general competence in areas that seem related. Usually that estimate is wildly overinflated."[71] Recent years' fixation on benchmarking has doubtless contributed to the misleading sense that LLMs' capabilities are rapidly moving toward those sufficient for autonomous deployment.

This issue of transferring to reality is often compounded by the advancements of LLMs along some measures. Thus far, we have concerned ourselves with the guaranteed performance of LLMs, which can be assessed through the accuracy and consistency of a model's outputs. However, guaranteed performance is not the sum total of safety-critical preparedness.

The ability to make sense of system's outputs in a given context (to *interpret* them), to audit the processes that lead to its outputs (to *explain* them), and to act on the system's internal components such that its outputs can be steered (to *intervene* on it) — are each relevant and potentially necessary.

The U.S. National Institute of Standards and Technology's (NIST) AI Risk Management Framework 1.0 defines some of these attributes separately:

> **Interpretability:** The *meaning* of a system's output in the context of those outputs' designs functional purposes.
>
> **Explainability:** A *representation* of the system's underlying mechanisms is available to the user.[72]

On the ability to intervene, Acharya and Song characterize it as follows:

> **Intervenability:** The ability to actively modify or interact with a model's internal mechanisms or representations to influence its outputs in a controlled and predictable manner. Effective intervention moves from the understanding of a model's internal state to acting on this understanding in a targeted fashion.[73]

All three matter for safety-critical deployments of AI systems, as they afford the model transparency during operations, which in turn allows for redress in relation to incorrect or negatively impactful outputs.

The rise of "reasoning"[74] models that output chains of thought — streams of text preceding a model's final answer that resemble verbalized human thought processes — are especially confounding with respect to interpretability, improving on benchmark scores without the requisite transparency. Indeed, it is unclear that these streams of text are usefully seen as *meaningful content* by the user, as they may not be faithful to the model's final answer.

[70] Vendrow, J., Vendrow, E., Beery, S., & Madry, A. (2025). "Do large language model benchmarks test reliability?" ArXiv, 10, https://arxiv.org/abs/2502.03461.

[71] See, Rodney Brooks' three laws of AI, here: https://rodneybrooks.com/rodney-brooks-three-laws-of-artificial-intelligence/.

[72] National Institute of Standards and Technology. (2023). "Artificial intelligence risk management framework (AI RMF 1.0)." https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, 16-17.

[73] Acharya, K. & Song, H. (2025). "A comprehensive review of neuro-symbolic AI for robustness, uncertainty quantification, and intervenability." Arabian Journal for Science and Engineering, 11. https://doi.org/10.1007/s13369-025-10887-3.

[74] See OpenAI's announcement here: https://openai.com/index/introducing-o3-and-o4-mini/.

Part of the reasoning models' appeal is their ability to break down problems into chains of thought that look startingly similar to a human's verbalized thought process (as though the model is talking to itself, as a person does, to work out a problem).[75] This technique is meant to make the model more adept at generating possible reasoning trajectories — possible chains of thought beginning with a problem and ending with a solution — and weigh them accordingly. The basic motivation is that these models capture a reasoning process closely resembling human reasoning, thereby making them more capable in ways relevant to the automation of human cognitive activities.[76]

However, these models' human-like outputs may not be what they seem. The same ASU research group that studied the planning abilities of LLMs conducted a study in May 2025 to assess the *relationship* between the accuracy of these chains of thought and the model's final (actual) answer to a problem.[77] The researchers tested this in the following way: They trained three transformer models on the solution-finding processes of a separate path-finding algorithm (the algorithm's steps in solving in solving a problem). Then, these models were each tasked with formulating plans to navigate a grid through permissible actions — up, down, left, and right.[78] Bizarrely, the model with the best performance — even showing slightly elevated performance beyond its training data — was the one whose training data (the algorithm's problem-solving steps) were *deliberately corrupted* by the ASU researchers.[79]

The implication of this finding is that chains of thought — though they strongly resemble a coherent, human-like thought process in verbalized form — do not hold the *meaningful content* we would ordinarily attribute to such text. In other words: Chains of thought are not reliably predictive of final answers. Safety-critical deployments that feed the human unreliable chains of thought are therefore limited in that these outputs cannot be meaningfully understood within the relevant context; their interpretation is essentially *ad hoc* and subjective in ways that depart from safety-critical standards.

This is not to suggest that constructing AI models that output chains of thought are *inherently* unreliable; it merely suggests that today's models are engaged in something not worthy of the term. The term "chains of thought" makes a promise: that models are engaged in verbalized reasoning processes that resemble that of a diligent human's workflow. This promise is akin to roboticist Rodney Brooks' dictum: that the *form* of a robot makes a *promise* about that robot's *capabilities*. If the robot has a humanoid form, it makes the promise that it can do all the things a human can do. So, too, for "Chains-of-Thought" and the "reasoning" models they are primarily instantiated in today. This promise is currently unmet.

Objections and Responses

The policy recommendations below aim to resolve these problems through federal policymaking that supports burgeoning hybrid AI research in the neuro-symbolic paradigm. However, ML is an unusual field, undergoing an even more unusual political moment. Some objections should be given fair consideration.

Objection 1: Complex ML will always be error prone, and we must learn to live with this. ML's incorporation into safety-critical domains may require that models be segmented and boxed off from other components in the system to prevent their errors from spreading and accumulating. Nevertheless,

[75] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). "Chain-of-Thought prompting elicits reasoning in Large Language Models." ArXiv, 2. https://arxiv.org/abs/2201.11903v6.

[76] Kambhampati, S., Stechly, K., & Valmeekam, K. (2025). "(How) Do reasoning models reason?" ArXiv, 2-3, https://arxiv.org/abs/2504.09762v1.

[77] Valmeekam, K., Stechly, K., Palod, V., Gundawar, A., & Kambhampati, S. (2025). "Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens." ArXiv, 1-16, https://arxiv.org/abs/2505.13775.

[78] Ibid., 3.

[79] Ibid., 7.

Researchers work at the research and development center of Sigma Geography, a geographic sciences multimodal large language model (LLM) in September 2024. (Li Xin / Xinhua via Getty Images)

aiming for the five nines of reliability or more (at least 99.999% accuracy) is unrealistic.

Response: It might be true that ML will always be error prone. This merely reinforces the need to foster the development of a research agenda for neuro-symbolic AI in high-stakes applications.

Moreover, it would be a nonstarter to incorporate into safety-critical systems ML subcomponents that have error levels higher than is traditionally expected, as subcomponents of the system must — at a minimum — interface with one another before delivering results to the end user, who needs to *trust* its outputs. An unacceptably error-prone ML subcomponent threatens the integrity of the entire system. If the ML subcomponent were to be segmented off to such a degree that its outputs are not critical for functionality, then the project of leveraging state-of-the-art ML's capabilities is sharply diminished.

Objection 2: Shifting resources away from ML research and development toward an immature and unproven research program like neuro-symbolic AI is unwarranted. Although complex ML models today experience shortcomings that make them unsuitable for a variety of safety-critical domains, the trendlines of research and development since at least 2022 indicate they are on the path to meeting these standards in due course.

Response: The premise is false. Neuro-symbolic AI funding does not require a shift, or at least not a significant one, away from ML R&D. The recommendations are indeed modest in their funding proposals and are meant to complement the private sector's overwhelming deployment of capital in ML R&D for commercial purposes.

To be sure, neuro-symbolic AI could fail to make the inroads necessary to achieve the ends of safety-critical integration. Yet, research paradigms cannot be judged this way, as if we were to suggest that funding for ML in its early days should have choked off, preventing researchers critical to the development artificial neural networks like John Hopfield from accessing National Science Foundation funding in the 1980s.[80] Indeed, neuro-symbolic AI today looks something like ML before its heyday: a long shot based on the enthusiasm of a small subset of AI researchers, hardly more promising than the more established symbolic AI of the day. Yet its own peak may be reached in the years ahead.

There is also wisdom in hedging one's bets, even where progress is indisputable. The transformer that underlies LLMs rules the day, at least for now. Even if matters like interpretability were resolved for these models, such techniques could prove ineffective should another architecture supplant it. This point is made nicely by researcher Grace Kind, who argues that — although it is, from her view, unlikely — a non-transformer architecture could return interpretability research to "square one."[81]

Finally, there is a possibility — however unlikely it seems — that testing of models will continue to occur in the real world through premature deployments but will not lead, either in due time or at all, to more reliable systems. It is possible that standards never catch up to reality because the transition was never pushed. In this scenario, having maintained roughly the current course, lives will have been harmed and lost because of these premature deployments without the gains necessary to reliably prevent this in the future.[82]

This somewhat resembles the current situation with self-driving vehicles. As Cummings observes:

> *In self-driving pilot operations around the country, state agencies have determined that the risk of public injury or death is worth the deployment of experimental Agentic AI. This effectively means that the public bears the brunt of testing in order to mature a technology*.[83]

Although extensive work is being conducted on the safety and real-world suitability of self-driving,[84] such deployments continue to risk and cause harm, as well as rely on humans where vehicle autonomy falls short.[85]

A future in which this trajectory remains as it is should be aggressively hedged against.

Objection 3: The United States is in a competition with China to rapidly develop and deploy powerful AI models, for which the current trajectory of deep learning appears best suited. Shifting state and commercial attention to narrow use cases is a distraction from this higher priority. The focus on safety-critical applications is too narrow and distracts from the widespread applicability that ML in its state-of-the-art, general-purpose forms has found.

Response: Although much policy analysis has focused on the pursuit of

[80] National Science Foundation News. (2024, October 8). "NSF congratulates laureates of the 2024 Novel Prize in physics." https://www.nsf.gov/news/nsf-congratulates-laureates-2024-nobel-prize-physics.

[81] Kind, G. (2025, November 22). "Will LLM alignment scale to general AI alignment?" https://gracekind.net/blog/llmalignment/.

[82] A parallel can be drawn here between premature deployments in safety-critical domains and what Sarah Fox and Samantha Shorey call *augmentation washing* — the reorganization of labor such that *human* work is structured around the *shortcomings* of machines, rather than serving the ostensible purpose of human augmentation. See, Fox, Sarah E. & Shorey, Samantha. (2025, December 22). "How augmentation-washing hides labor automation." Tech Policy Press. https://www.techpolicy.press/how-augmentationwashing-hides-labor-automation/.

[83] Cummings, M. (2025). "Prohibiting generative AI in any form of weapon control." NeurIPS 2025 Position Paper, 10, https://openreview.net/forum?id=uEY7kQsiZz&referrer=%5Bthe%20profile%20of%20Mary%20Cummings%5D(%2Fprofile%3Fid%3D~Mary_Cummings1).

[84] Dong, X. et al. (2025). "Why autonomous vehicles are not ready yet: A multi-disciplinary review of problems, attempted solutions, and future directions." Journal of Field Robotics, 1-88. https://doi.org/10.1002/rob.70108.

[85] Somewhat akin to augmentation washing described in footnote 87, ostensibly "autonomous" vehicles frequently require human assistance. Waymo, for example, uses an app called "Honk" where individuals are paid one-off sums of money ($20 or more) to close Waymo vehicle doors that are left open so that the vehicle can continue moving. See, Bonos, Lisa. (2025, December 25). "When robot taxis get stuck, a secret army of humans comes to the rescue." The Washington Post. https://www.washingtonpost.com/technology/2025/12/25/waymo-robots-human-work/.

ever-more capable AI models, such pursuits have been more short-sighted than appreciated. It is the integration of powerful yet trustworthy models in systems that impact the most sensitive domains of human life that signify the fullest notion of progress. To use another analogy, a world in which air travel is possible but where only four out of five flights avoid crashing is unimaginable. The capability of flight was instead made safe for humans, impacting it in ways difficult to imagine otherwise. ML must likewise be made safe for the world.

The focus on safety-critical domains is the rightful high bar for this technology to clear. The United States risks misleading itself into imagining the pursuit of AI capabilities for their own sake as equivalent to the advancement of society itself. But only trustworthy technologies that are built with respect to their limitations and the dignity of those who use them achieve this end. Insofar as competition with China is a factor, this point stands no less firm in its wake.

"The focus on safety-critical domains is the rightful high bar for this technology to clear. The United States risks misleading itself into imagining the pursuit of AI capabilities for their own sake as equivalent to the advancement of society itself."

Objection 4: AI is fundamentally different from other technologies. Unlike other powerful technologies that have become "normal technologies" over time, AI is in fact not a normal technology.[86] We must approach it on different terms.

Response: Cummings' point that the development of generative AI models was short-circuited by a rapid transition from laboratory development to commercialization is worth recalling here: That these models appear fundamentally different from other technologies is a symptom of our lack of understanding of them. Time was not given for a theory of how these models generalize to develop, for interpretability tools to be available to end-users, or to find ways to minimize their probability of harmful outputs and severity therein. Deployment has outpaced the *kinds of progress* necessary for higher-impact usefulness.

State-of-the-art ML models are not normal technologies. Yet this is to be expected. The aim of the recommendations is to make them so.

[86] Policy analyst Peter Wildeford has made various informal comments to this effect. See, e.g., https://substack.com/@peterwildeford/note/c-154147347.

RECOMMENDATION

The ML technical agenda must be shifted to prepare powerful, complex models for safety-critical domains. It is insufficient to seek organizational changes that accommodate unreliable technologies. Thus, several federal policy changes should be undertaken in short order to this end, implicating research and development, evaluation standards, and public-private dynamics.

1 INSTRUCT RESEARCH AND DEVELOPMENT

The National Artificial Intelligence Initiative Office (NAIIO), together with the Networking and Information Technology Research and Development (NITRD) Subcommittee, should instruct the National Artificial Intelligence Research & Development Interagency Working Group to coordinate investments in safety-critical, application-specific neuro-symbolic AI Research. These investments should be valued at $200 million in the first year of such coordination. Investments should increase annually over the next four years, totaling approximately $1.5 billion over the five-year period. This sum does not include private-sector contributions in public-private partnerships.

Explanation: By using its role as a coordinating body, the NAIIO props up burgeoning efforts currently restricted to the minority of the private sector's AI R&D, hedging the U.S. government's bets on the ability of a primarily or strictly ML paradigm to resolve critical limitations in high-stakes domains.

2 INFORM AND HELP GUIDE INVESTMENT

The NIST Center for AI Standards and Innovation (CAISI) should inform — but not exclusively guide — the NAIIO's investment coordination by convening a working group that establishes a new evaluation metrics research agenda for safety-critical neuro-symbolic systems. This working group is to be made up of researchers from both industry and academia, particularly those whose work has operated on the peripheries of commercial AI R&D. It is not necessary that researchers identify their work as "neuro-symbolic;" it is more important that their approach is tailored to safety- or mission-critical domains and avoids excessive reliance on data-driven techniques. The working group should meet over the course of three to six months, with instructions to achieve certain goals:

- Identify the applications for which neuro-symbolic AI research is already being done and which application areas are most suitable for the near-term exploration of novel evaluation metrics.[87]
- Identify a candidate list of techniques already being explored in the relevant application areas for which funding is likely to have the highest impact. If none exist for the target domains, the working group should identify research most relevant to these foundational matters.
- With applications and potential techniques identified, determine whether new datasets must be created to aid in the modeling of relevant real-world dynamics for these application areas. In either case, the NAIIO will use this information during its investment coordination.
- Provide final recommendations to the NAIIO.

Explanation: Evaluation metrics for neuro-symbolic models in safety-critical

[87] See, Renkhoff, J., Feng, K., Meier-Deornberg, M., Velasquez, A., & Song, H. H. (2024). "A survey on verification and validation, testing and evaluation of neurosymbolic artificial intelligence." ArXiv, 11-12. https://arxiv.org/abs/2401.03188 for related discussion.

domains will not look like the "benchmark maxing" of recent years, in which models are trained to improve their accuracy on static datasets of various problems. Rather, real-world dynamics are the target of safety-critical evaluation.

3 DETERMINE IF AUTONOMY READINESS LEVELS ARE WARRANTED

A separate working group of similar composition, also convened by NIST's CAISI, should determine whether Autonomy Readiness Levels (ARLs) for non-defense, safety-critical applications are warranted at the current stage of development. These are evaluation metrics specifically for models expected to perform autonomously over a potentially shifting range of circumstances in the execution of a human-given input. AI "agents" may be considered a part of this target group. This working group should meet several times within the span of one month.

These ARLs are inspired by the U.S. DoD's Technology Readiness Levels. Defense-specific efforts are underway through at least two DARPA projects: ASIMOV and Artificial Intelligence Quantified (AIQ).

DARPA Program Managers overseeing the above or related efforts should be invited to the working group in their capacity as technical experts, but the group's efforts will not hinge on their participation.

The working group should have the following goals:

- Study the origin and development of the DoD's TRLs and explicate their underlying structure in abstract terms.
- Transfer the structure of this taxonomy to the development of commercial AI systems that operate autonomously in safety-critical domains and determine if, at the current stage of development, ARLs can be fruitfully applied to autonomous systems, like agents.
- Provide recommendations to the NAIIO on whether this research is worth funding.

Explanation: Sensing that the time is ripe for a robust conceptualization of autonomous systems, commercial efforts should be undertaken akin to DARPA's, acknowledging the increasing presence of autonomous systems in everyday life will require the utmost scrutiny of their integration into safety-critical domains.

4 NATIONAL SCIENCE FOUNDATION SHOULD ESTABLISH AI RESEARCH INSTITUTE

The NSF should parallel the NAIIO's efforts by establishing a National AI Research Institute dedicated to a specific application area of neuro-symbolic AI. An initial investment worth between $80 million to $100 million over five years, subject to changes, should be made. This investment is $60 to $80 million above what these Institutes are typically given for their initial five-year periods.

Explanation: The NSF's establishment of a new AI Research Institute recognizes that there is a plausible outcome in which ML as it is currently constituted in broadest formulation cannot meet the needs of safety-critical domains while retaining the capabilities earned by its state-of-the-art models. A knock-on effect will be to signal that, where private actors are willing, the U.S. government may meet them part of the way in the deployment of capital to serve this end, given the National AI Research Institutes' emphasis on public-private partnerships and other forms of collaboration.[88]

[88] On this collaborative structure, see, Donlon, J. J. (2024). "The National Artificial Intelligence Research Institutes program and its significance to a prosperous future." AI Magazine, 45(1): 6-14. https://doi.org/10.1002/aaai.12153.

Conclusion

The deep learning revolution, such as it is, should come to an end. This is not for lack of progress; capabilities have been gained and progress has continued along certain measures. Rather, it is because all revolutions must eventually come to an end and new orders must take their places, putting aside or better channeling the excesses which fueled their rise. The capabilities being gained by the dominant paradigm in AI today are not what is needed for its models to integrate into safety-critical domains. Individuals and social institutions should not lower their standards to accommodate unreliable technology, and current priorities in ML research show little concern for fixing this problem.

The future of ML and AI in the highest-impact, safety-critical domains will not resemble the heavenly or nightmarish visions of "Artificial General Intelligence" or "Superintelligence." The technology's capabilities will, in the event of successful integration, respect both its limitations and the dignity of those interacting with it. The beauty and terror often seen in AI today is a reflection of the viewer's beliefs about humanity and the possibilities that something made in our image could replicate our best and worst traits. The instinct is understandable. But the course-correction from the deep learning revolution that is now needed is to understand that technology's beauty or terror comes from the standards to which we hold it. ML must change to serve them.

> "The capabilities being gained by the dominant paradigm in AI today are not what is needed for its models to integrate into safety-critical domains. Individuals and social institutions should not lower their standards to accommodate unreliable technology, and current priorities in ML research show little concern for fixing this problem."

The views expressed in this article are those of the author and not an official policy or position of New Lines Institute.

AUTHOR

Vincent J. Carchidi is a Defense Industry Analyst at *Forecast International*, where his work focuses on defense technologies, with a particular interest in autonomy and the transition of emerging technologies to suitability for mission-critical domains. He has a background in U.S. technology policy, contributing to U.S.-Middle East technology policy, with particular interests in artificial intelligence and critical infrastructure. He maintains an academic background in cognitive science and philosophy of mind that informs his policy work.

Carchidi's opinions are entirely his own and do not reflect those of his employer.

NEW LINES INSTITUTE
FOR STRATEGY AND POLICY

Contact

For media inquiries, email
media@newlinesinstitute.org

To learn more about New Lines' publication process, email
submissions@newlinesinstitute.org

For other inquiries, send an email to
info@newlinesinstitute.org

A: 1660 L St. NW, Ste. 450
Washington, D.C., 20036

P: (202) 800-7302