The Alan Turing Institute

Centre for Emerging Technology and Security

RESEARCH REPORT

The Next Frontier: Security Implications of Future AI Paradigms

Ardi Janjeva, Sylvester Kaczmarek, Angus Shennan and Carolyn Ashurst

April 2026

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at the Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to strengthen UK security through pioneering research on emerging technologies. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by the Alan Turing Institute's Defence and National Security grand challenge. All views expressed in this report are those of the authors, and do not necessarily represent the views of the Alan Turing Institute or any other organisation.

Acknowledgements

The authors would like to thank all those who contributed their time to participate in research interviews and provide survey responses for this project. The authors would also like to thank the following contributors for their valuable feedback on an earlier draft of this report: Andrew Duncan, Jason McEwen and Ture Hinrichsen.

Cite this work as: Ardi Janjeva, Sylvester Kaczmarek, Angus Shennan and Carolyn Ashurst, "The Next Frontier: Security Implications of Future AI Paradigms," *CETaS Research Reports* (April 2026).

Executive Summary

The current paradigm of frontier AI – driven by massive compute and transformer-based generative models – continues to yield significant value. However, as compute scaling faces potentially diminishing returns, future AI progress is unlikely to rely on a single, isolated breakthrough. The next wave of capability improvements will most likely emerge from a complex mix of extensions, hybrids and specialised pathways combined with existing models.

This report maps the emergence of novel paradigms across alternative architectures, data sources, learning methods and hardware substrates. While some approaches, such as agentic scaffolds, readily complement existing systems, others – like neuro-symbolic reasoning, world models and neuromorphic computing – could represent more divergent pathways. Because frontier-scale generative models are not a universal solution for every deployment context, many of the approaches surveyed will also lead to developments in particular domain areas.

This report provides a basis to help policymakers anticipate and respond to technological surprise. For the UK, maintaining a competitive edge in AI requires strategic agility – the ability to respond and pivot in response to rapid developments. To achieve this, the report advocates for action across three priority policy areas: building a deep skills base; investing in supporting infrastructure; and embracing fast-following and application at scale.

Key findings

The existing paradigm

- The current AI paradigm is not yet exhausted, but its **next gains are increasingly likely to come from a broad, complementary package of methods**, including developments in post-training, reinforcement learning (RL), inference-time compute, tool use, and system design, rather than from scaling pre-training alone. The drive to attain new sources of data, first through non-text media for pre-training, and then through post-training avenues like RL, has partly mitigated concerns around an AI data bottleneck.

- Future AI progress is likely to come from a **mix of extensions, hybrids and specialised pathways** rather than one breakthrough in isolation. However, because

current approaches have been scaled and optimised so vigorously, there is a **risk that the potential of brand-new approaches may be masked**.

- In most plausible trajectories, **scaling compute will remain a differentiating factor**. Without having access to large-scale compute, the inability to scale research breakthroughs would likely see them more rapidly absorbed elsewhere.
- There are many research challenges that need solving within the current paradigm. Frontier labs tend to have the scale and resources needed to work on many of these, with some also actively researching alternative approaches alongside the existing paradigm.
- The **automation of AI R&D** could also be pivotal in both **accelerating individual research paths and determining the number of paths that can be pursued simultaneously**.

Novel paradigms

- There are **other research approaches** besides the existing AI paradigm that will likely lead to significant capability developments. Some, such as agentic systems and structured retrieval, are already being deployed in commercial systems and should be understood as active extensions of the current frontier paradigm. Others remain more exploratory, with greater uncertainty around whether they will mature and scale.
- **Nearer-term developments** include those that extend the current paradigm to afford additional capabilities or solve practical constraints. These include **agentic systems, structured retrieval, continual learning, efficient deployment methods, and some alternative hardware pathways**.
- More **medium-term or uncertain areas include larger-scale world models, more developed embodied learning loops, broader deployment of neuromorphic systems** beyond niche contexts, and **thermodynamic computing** beyond early prototypes.
- **Longer-range or speculative areas include quantum machine learning**, which currently lacks decisive evidence as to whether it will ultimately emerge as a viable solution.

- **The strategic significance of alternative paradigms is often derived from usability, cost and governability** as much as from raw capability alone. Several of the most important developments and uses of agentic systems, lean models, neuromorphic computing and alternative hardware relate to systems that are best adapted to specific deployment contexts.

- **The ability to test, monitor, govern and deploy new paradigms with confidence** is vital. Models acting through tools and external data sources create a wider surface for error and misuse; a hardware pathway with weak tooling may struggle to move beyond promising demonstrations; and a continually updated model may create new integrity and rollback problems.

The national security and strategic landscape

- **Asymmetrical AI capabilities between adversaries resulting from more globally diverse technical approaches could exacerbate deterrence and competition dynamics** in military, economic and geopolitical contexts. Asymmetries could lower the threshold for conflict for leaders in military applications of AI.

- **UK Government anticipatory capabilities have improved** since the popularisation of large language models (LLMs), **but better anticipation of technology trajectories is not sufficient for readiness**.

- As AI becomes more embedded in critical infrastructure, the stakes associated with maintaining access to frontier AI systems will rise. To hedge against technological uncertainty, consideration must be given to the infrastructure required under different AI trajectories.

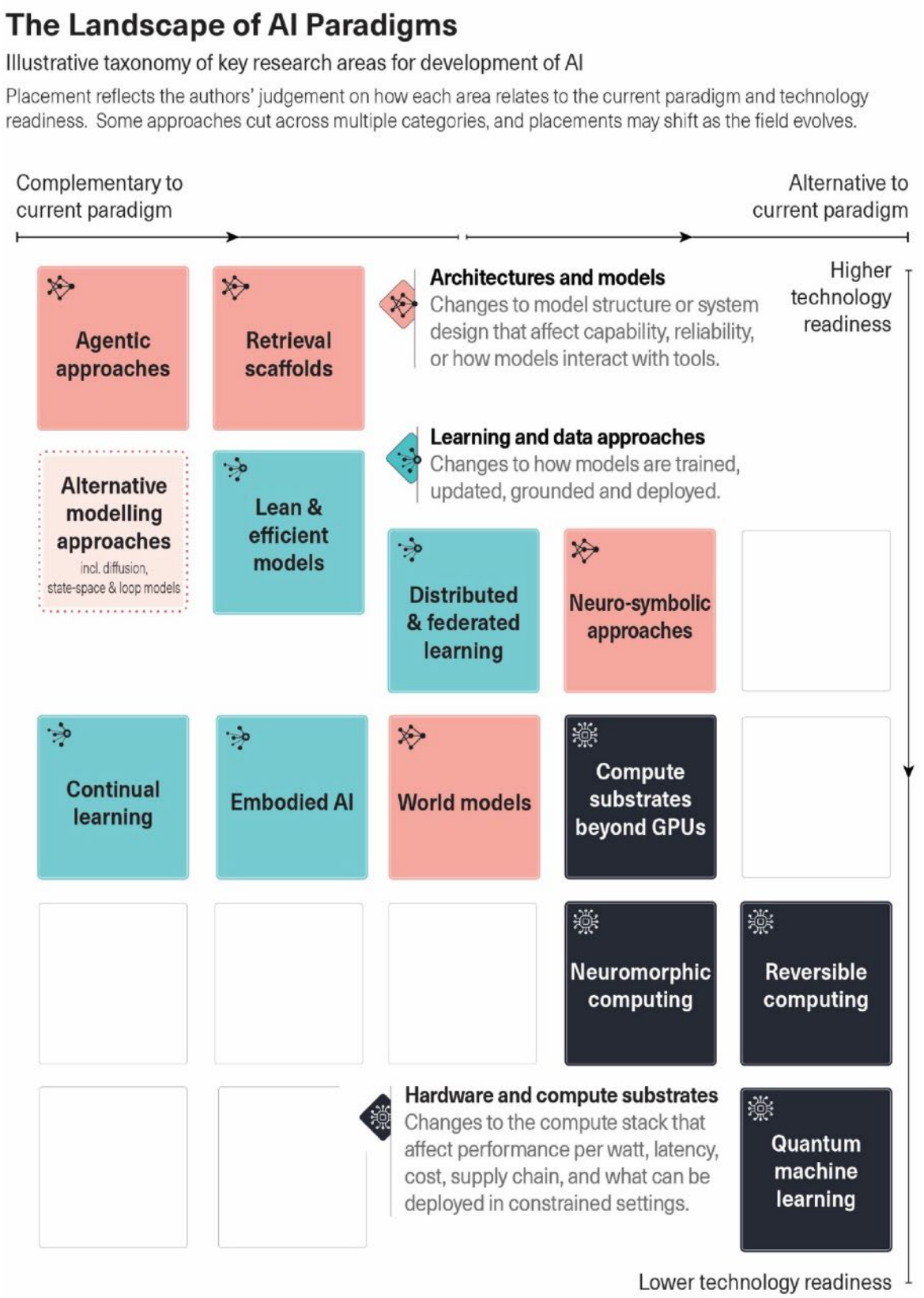
- The UK can adopt strategies to improve its **ability to respond to a range of AI futures**. These include:

    - **Building a deep skills base** – including increasing i) expertise in developing and maintaining large-scale models, ii) transferable technical skills, and iii) specialised skills needed for different paradigms discussed in this report.

    - **Investing in supporting infrastructure** – including a fine-tuning AI stack that allows adaptation to domain-specific data to build specialised tools, a secure sandbox for developing and testing next-generation agentic AI capabilities before deployment, and shared access to testbeds for emerging hardware pathways.

- **Fast-following and application at scale** – focusing on the application of AI systems (including specialised lean models) to high-value sectors that convert model performance into economic and strategic value, while reducing friction in government adoption.

Figure 1. Overview of AI paradigms discussed in this report



1. Introduction

Over the last decade, the most significant developments in AI have been largely driven by a single, dominant architectural approach. This report uses the term 'current paradigm' to refer to today's frontier AI systems. These typically comprise frontier-scale, transformer-based, multimodal generative models, which rely on massive pre-training coupled with post-training methods (such as RL). The current paradigm also captures how these models are deployed into larger computational systems (such as Claude and Gemini) and include elements to support additional functionality, guardrails and user interface.

While this paradigm continues to yield impressive capabilities, the established strategy of continually scaling compute is encountering physical and economic boundaries.[1] For example, projections from Epoch AI indicate that the supply of high-quality, publicly available human text is approaching its limit.[2] As scaling laws dictate that exponential increases in compute yield log-linear capability gains,[3] sustaining the current trajectory requires increasingly vast resources. With the next generation of frontier AI data centres requiring hundreds of megawatts – and future clusters projecting gigawatt-scale power demands – the thresholds for foundational pre-training are only available to a select few actors.[4]

The current paradigm is effective at extracting insights from large unstructured datasets, automating high-volume administrative workflows, and accelerating software engineering through automated code generation. But this does not mean that today's frontier systems are optimal for every deployment context. Current systems can fail unpredictably on complex, out-of-distribution tasks, lacking the verifiable outputs required for decision-making in domains like national security.

Recognising these (and other) limitations, research is progressing in numerous directions. Much of this innovation complements rather than replaces the current approach, suggesting that important gains can still be achieved within today's paradigm. Simultaneously, alternative approaches are showing promise in narrow domains, ranging

[1] Tom Davidson, Rose Hadshar and William MacAskill, "How Far Can AI Progress Before Hitting Effective Physical Limits?," *Forethought*, 17 March 2025, https://www.forethought.org/research/how-far-can-ai-progress-before-hitting-effective-physical-limits.

[2] Pablo Villalobos et al., *Will we run out of data? Limits of LLM scaling based on human-generated data* (Epoch AI: June 2024), https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

[3] Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* (Google DeepMind: March 2022), https://arxiv.org/abs/2203.15556.

[4] Venkat Somala and Ben Cottier, "Build times for gigawatt-scale data centers can be 2 years or less," *Epoch AI*, 10 November 2025, https://epoch.ai/data-insights/data-centers-buildout-speeds.

from neuro-symbolic systems and world models to hardware shifts like neuromorphic computing. Many alternative approaches are intended to solve different problems, rather than being a like-for-like successor. We are likely to see a hybrid future where different approaches play different roles.

For the UK, maintaining a competitive edge in this diverse landscape requires a deep understanding of current AI capabilities (and emergent risks and opportunities) and strategic agility. To achieve this, the report advocates for action across three priority policy areas: building a deep, adaptable skills base; investing in versatile supporting infrastructure; and embracing a posture of fast-following and application at scale.

The report first unpacks the existing paradigm (Section 2), assessing how much untapped potential remains, the role of scaling for continued progress, the potential for automation of AI R&D, and the technical challenges that still need solving. It then evaluates a series of novel approaches against these challenges through a taxonomy focusing on architectures and models, learning and data paradigms, and hardware and compute substrates (Section 3). Finally, it outlines what the UK should do in the face of uncertain AI trajectories (Section 4).

1.1 Methodology

This study was conducted between December 2025 and March 2026. Data collection involved the following core research activities:

- Review of academic and grey literature.
- Research interviews with 25 experts across the UK Government, industry and academia.
- A small-scale survey with N = 25 respondents targeted at AI experts in academia and industry. Dissemination of the survey was primarily conducted through the Turing University Network, which has a membership of 65 universities from across the UK. Individual liaisons representing each university were asked to identify a selection of experts within their institutions. Invitations to complete the survey were also extended to Turing AI Fellows.
- Analysis of publications from the OpenAlex repository of academic works, using keyword search and topic modelling to identify relevant publications. Keywords were selected through expert consultation to cover variation in terminology. Data is current as of 11 March 2026.

Limitations of this study include:

1) Some important frontier developments are only partially visible because they sit within private companies rather than published papers.

2) The literature is uneven and noisy across many of the paradigms analysed in this report due to rapid changes in the field, and different approaches to publishing across different areas.

3) Taxonomy boundaries are blurred and some novel research approaches could plausibly sit in more than one category.

4) Interview data collected for this project is highly valuable for providing specific insights. While a breadth of expertise and professional backgrounds was targeted, views will not be representative of the whole field.

5) Due to the small scale of the survey, the sample should not be interpreted as representative of any group or profession; general trends are instead presented for additional nuance to complement the other methodological components. Future research could seek to establish a comparison with an industry- or government-focused cohort.

2. The Existing Paradigm

2.1 Untapped potential or hitting a wall?

There is vigorous debate between those who think that the current AI paradigm is approaching (or at) the limits of its capabilities, and those who think that we are still scratching the surface of what can be achieved.[5] Two-thirds of the limited survey sample for this research saw a plateau in frontier model capability progress by 2028 as either likely or very likely.

The assessment of this report is nuanced. The current paradigm is not exhausted, but its next gains are increasingly likely to come from a broader package of methods, including developments in post-training, RL, inference-time compute, tool use, and system design, rather than from scaling pre-training alone.[6] We can expect significant capability increase based on a *combination* of different architectures involving the current approach – this may look quite different to the typical models we see today. Survey data saw 'RL and preference optimisation' as the most selected factor to extend the dominance of transformer-based systems.

The progression seen with test-time compute since mid-to-late 2024 represents an example of this layering. This includes developments in *reasoning*: methods that provide granular feedback on intermediate reasoning steps, and allow a model to explore and backtrack through multiple potential solution paths.[7] Additionally, with systems like OpenAI's o1/o3 and DeepSeek's R1, improved *model scaffolding* has been shown to enhance model performance (by using hidden deliberation loops and internal self-correction) and increase compute efficiency.[8]

In recent months, we have seen steady progress against some important limitations, such as performance on long tasks, error rates, hallucinations, and the ability to navigate digital

[5] Cade Metz, "An AI Pioneer Warns the Tech 'Herd' Is Marching Into a Dead End," *New York Times*, 26 January 2026, https://www.nytimes.com/2026/01/26/technology/an-ai-pioneer-warns-the-tech-herd-is-marching-into-a-dead-end.html; The Economist, "Nvidia's boss dismisses fears that AI has hit a wall," 21 November 2024, https://www.economist.com/business/2024/11/21/nvidias-boss-dismisses-fears-that-ai-has-hit-a-wall; Ethan Mollick, "The Shape of the Thing," *One Useful Thing*, 12 March 2026, https://www.oneusefulthing.org/p/the-shape-of-the-thing.
[6] Author interview with academic participant, 15 January 2026; Author interview with academic participant, 19 January 2026.
[7] Charlie Snell et al., *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters* (Google DeepMind: August 2024), https://arxiv.org/pdf/2408.03314.
[8] AI Security Institute, *Frontier AI Trends Report* (December 2025), https://aisi.s3.eu-west-2.amazonaws.com/Frontier+AI+Trends+Report+-+AI+Security+Institute.pdf.

environments.[9] However, there are other components like spatial reasoning and understanding, planning and optimising that seem likely to require new ingredients.[10]

2.1.1 Circumventing the data bottleneck: the role of post-training

The drive to attain new sources of data, first through non-text media for pre-training (such as image, video and audio) and then through post-training methods like RL, has partly mitigated concerns around an AI data bottleneck. By using RL to teach models how to better self-correct, the industry has unlocked a new axis for scaling: improving reasoning and other capabilities by scaling compute at inference time, rather than just ingesting more data during pre-training. Coding has proven to be a strong domain for learning in this way due to its objective and verifiable nature, and there is plenty of work underway trying to apply this approach to other domains.[11]

For domains that lack an objective, automated way to verify success, scaling RL presents a bigger challenge. More broadly, there is concern around the extent to which using model-generated data for training could cause 'model collapse',[12] and the possibility that RL models are overfitting to specific benchmarks and reward signals.[13] Furthermore, the necessary data collection can be slow and capital-intensive, with data quality needing to be prioritised over quantity.[14]

Notably, however, this focus on data quality may lead to more differentiation between models: whereas during the pre-training era, the scale of data was paramount, with RL training, AI developers need to think harder about the variety of RL environments used.[15]

---

[9] Author interview with government participant, 26 January 2026; OpenAI, "Introducing OpenAI o3 and 04-mini," 16 April 2025, https://openai.com/index/introducing-o3-and-o4-mini.

[10] Author interview with industry participant, 10 February 2026.

[11] Author interview with industry participant, 21 January 2026; Subhadip Mitra, "RLVR Beyond Math and Code: The Verifier Problem Nobody Has Solved," 18 January 2026, https://subhadipmitra.com/blog/2026/rlvr-beyond-math-code.

[12] This is caused by iterative training on unrefined synthetic data, leading to a gradual loss of representativeness; Matthias Gerstgrasser et al., *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data* (Stanford University; Constellation; University of Maryland; MIT: April 2024), https://arxiv.org/html/2404.01413v2; Kareem Amin et al., *Escaping Collapse: The Strength of Weak Data for Large Language Model Training* (Google Research; University of Southern California: November 2025), https://arxiv.org/abs/2502.08924.

[13] Dwarkesh Patel, "Ilya Sutskever – We're moving from the age of scaling to the age of research," *Dwarkesh Podcast*, 25 November 2025, https://www.dwarkesh.com/p/ilya-sutskever-2.

[14] Author interview with industry participant, 21 January 2026; Author interview with government participant, 26 January 2026.

[15] Patel (2025); Ziqian Zhong, Aditi Raghunathan and Nicholas Carlini, *ImpossibleBench: Measuring LLMs' Propensity of Exploiting Test Cases* (Carnegie Mellon University; Anthropic: October 2025), https://arxiv.org/abs/2510.20270; Darshan Deshpande, Anand Kannappan and Rebecca Qian, *Benchmarking Reward Hack Detection in Code Environments via Contrastive Analysis* (Patronus AI: January 2026), https://arxiv.org/abs/2601.20103.

2.1.2 Realising the potential of existing capabilities

Irrespective of leaps in performance, many argue that more can be extracted from the current paradigm through deployment in new areas of the economy.[16] This could be enabled by the development of RL environments for new applications, such as the energy grid. Being at the forefront of model development is not the only factor in realising the potential of AI – being well placed to *apply* models in a wide range of sectors is imperative.

However, to achieve this, systems need to effectively navigate real-world settings and societal constraints and requirements – the knottier interactions with human structures and organisations. A model can score 95% on a coding benchmark, but in the real world, codebases are messy, human instructions can be vague, and legal and regulatory requirements apply.[17] This is encapsulated in METR's 2025 study of the developer productivity gap, which found that while AI agents could solve 90% of benchmarked coding issues, they slowed down experienced developers by 19%.[18] A follow-up study in late 2025 found suggestive but inconclusive evidence that this effect may be reversing as tools improve, though selection effects prevented reliable measurement.[19] AI for coding is one of the leading examples of economic value being extracted from commercial models, but even in this domain there remain challenges.

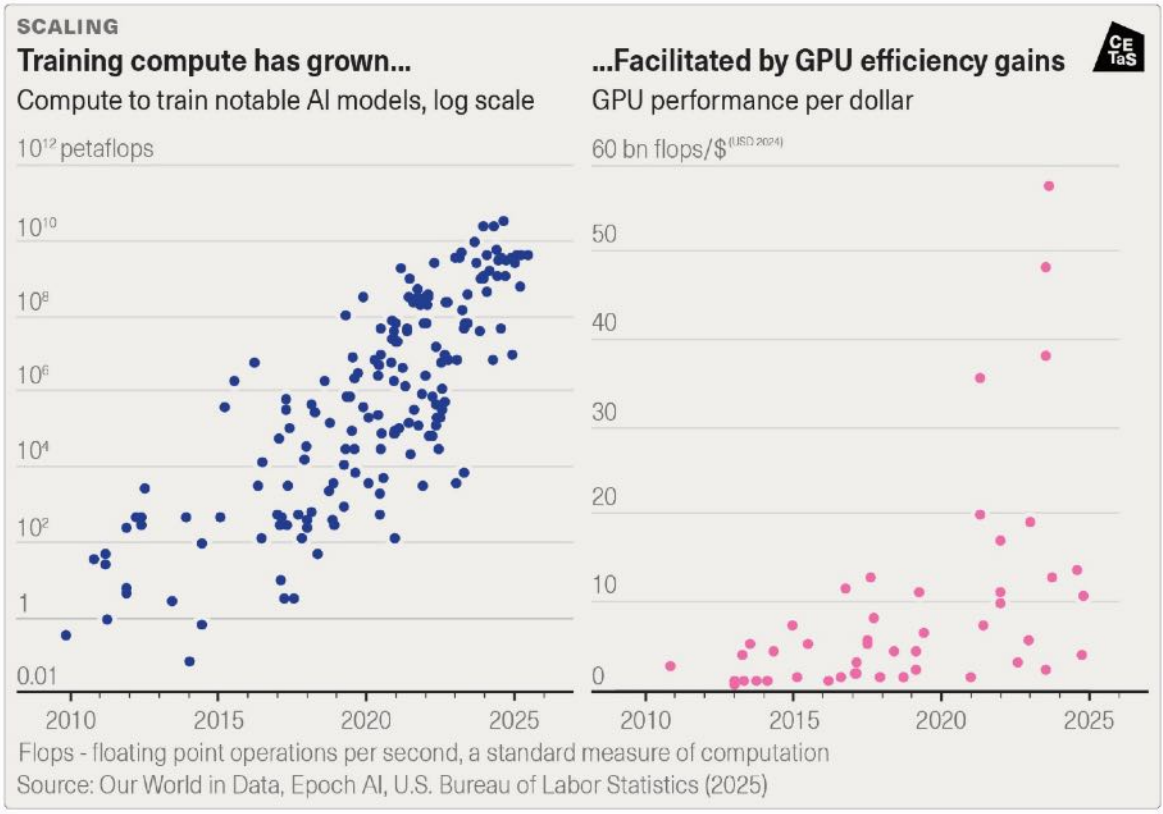
Finally, it is important not to overlook the surprise that might be generated in more subtle ways. In recent years, we have seen that step changes often come from a combination of incremental gains, or under-the-radar efficiencies that cause scaling curves to inflect.[20]

[16] Author interview with industry participant, 20 January 2026.
[17] Thomas Kwa et al., *Measuring AI Ability to Complete Long Tasks* (METR: March 2025), 39-40, https://arxiv.org/pdf/2503.14499v1.
[18] Joel Becker et al., *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity* (METR: July 2025), https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study.
[19] Joel Becker et al., *We are Changing our Developer Productivity Experiment Design* (METR: February 2026), https://metr.org/blog/2026-02-24-uplift-update.
[20] Nicholas Roberts et al., *Compute Optimal Scaling of Skills: Knowledge vs Reasoning* (Meta: March 2025), https://arxiv.org/html/2503.10061v1; Dan Busbridge et al., *Distillation Scaling Laws* (Apple: July 2025), https://arxiv.org/abs/2502.08606.

2.2 Importance of scaling

The term 'scaling laws'[21] refers to how increased computation power, training data, and model parameters has translated into increased performance in AI research.[22]

Figure 2. Increase in AI training compute and GPU efficiency



Frontier labs are investing in ever-larger data centres and energy infrastructure to scale compute. This is forcing a redesign of data centre infrastructure, shifting to high-voltage electrical grids and installing dedicated power racks that pump 1MW of electricity into single clusters of AI chips. Google recently announced plans to double capital expenditure in 2026,[23] while Meta announced in Q1 2026 that it will spend up to US$135 billion this year alone on AI infrastructure.[24] Forecasts of capital expenditure for major frontier labs have

[21] Jared Kaplan et al., *Scaling Laws for Neural Language Models* (OpenAI: January 2020), https://arxiv.org/abs/2001.08361; Hoffmann et al. (2022).
[22] Veronika Samborska, "Scaling up: how increasing inputs has made artificial intelligence more capable," *Our World in Data*, 20 January 2025, https://ourworldindata.org/scaling-up-ai.
[23] Stephen Morris, "Google set to double AI spending to $185bn after strong earnings," *Financial Times*, 5 February 2026, https://www.ft.com/content/22d97d8e-1101-4b1b-8a28-66054dfa363a.
[24] Meta Investor Relations, "Meta Reports Fourth Quarter and Full Year 2025 Results," 28 January 2026, https://investor.atmeta.com/investor-news/press-release-details/2026/Meta-Reports-Fourth-Quarter-and-Full-Year-2025-Results/default.aspx.

increased each year since 2024,[25] and worldwide, it is estimated that US$3 trillion will be spent on data centres from 2025 to 2029.[26]

2.2.1 The convenience of the 'bitter lesson'

In the early days of AI, researchers sought to improve model performance by leveraging domain-specific human knowledge about a problem. For example, embedding heuristics and structures in chess-playing engines. However, the so-called 'bitter lesson' of AI research has been that performance in the medium-long term has tended to improve by applying more general methods at higher compute.[27] Now, compute scaling is increasingly complementing the post-training phase (OpenAI prioritised scaling post-training with GPT-5, for example).[28] However, it is also argued that scaling generic models leads to another trap: training massive models involves scaling human knowledge, which includes human biases, misconceptions and flawed heuristics, and could miss the importance of anchoring to invariant properties of the world.[29]

'Scaling' is a powerful word because it presents companies with an equation for making incremental progress by allocating more compute.[30] However, in this environment, the potential of brand-new approaches may be masked, because they are compared to the current approach that has been scaled and optimised so vigorously. The environment also means that many pertinent questions are deprioritised, including how more could be done with more constrained resources, or how to realise the potential of hyper-personalised or domain-specific models.[31]

2.2.2 Finding the right balance

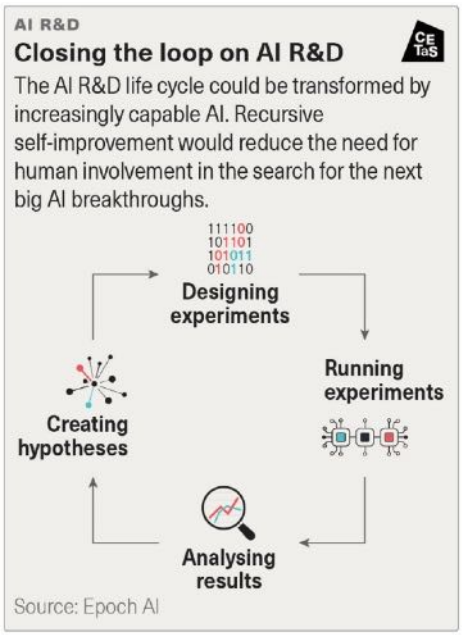
In the current landscape, with clusters already utilising over 100,000 graphics processing units (GPUs), simply scaling compute by 100x would yield only diminishing returns, partly

[25] Goldman Sachs, "Why AI Companies May Invest More than $500 Billion in 2026," 18 December 2025, https://www.goldmansachs.com/insights/articles/why-ai-companies-may-invest-more-than-500-billion-in-2026.
[26] Michael Dempsey, "What's the big deal about AI data centres?," *BBC News*, 23 September 2025, https://www.bbc.co.uk/news/articles/ckg2ldpl9leo.
[27] Richard Sutton, "The Bitter Lesson," 13 March 2019, https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf.
[28] Yafah Edelman et al., "Why GPT-5 used less training compute than GPT-4.5 (but GPT-6 probably won't)," *Epoch AI*, 26 September 2025, https://epoch.ai/gradient-updates/why-gpt5-used-less-training-compute-than-gpt45-but-gpt6-probably-wont.
[29] Jason McEwen, "A Nuanced Perspective on The Bitter Lesson," *Inductive AI Biases of Jason McEwen*, 4 February 2026, https://inductivebias.substack.com/p/a-nuanced-perspective-on-the-bitter.
[30] Author interview with academic participant, 2 February 2026.
[31] Ibid.

due to the exhaustion of high-quality human text data.[32] Moreover, such a 100x increase would require multiple gigawatts of power (several nuclear reactors' worth of energy).

Therefore, it is likely that we are approaching an equilibrium where more algorithmic innovation is needed, but the most impactful developments are likely to be those that scale with compute.[33] This is a common trend in the recent history of AI: the transformer was built on 8-64 GPUs,[34] and researchers had to prove that it could process data more efficiently than older models (like recurrent neural networks) to demonstrate that it should be scaled up.[35]

The insight that new paradigms will likely still be responsive to scaling compute is potentially sobering for countries looking towards new paradigms as a way of lessening dependencies on leading companies in the US. For example, US labs were well positioned to capitalise on test-time compute breakthroughs by scaling Monte Carlo tree search (MCTS) layers (a decades-old public-domain algorithm) across their entire stack within months.



2.3 Automation of AI R&D

Figure 3. Automating the AI R&D cycle

The automation of AI R&D (potentially leading to recursive self-improvement) refers to the use of advanced AI systems to train the next generation of AI systems. This is a stated goal for the major frontier labs and is increasingly seen as one of the most critical proxy measures for monitoring progress towards artificial general intelligence (AGI).[36] AI is already used to accelerate parts of the AI R&D workflow, especially coding tasks:[37] 50% of new code

[32] Villalobos et al. (2024).

[33] Author interview with industry participant, 20 January 2026.

[34] Rick Merritt, "What Is a Transformer Model?," *Nvidia*, 25 March 2022, https://blogs.nvidia.com/blog/what-is-a-transformer-model.

[35] Author interview with government participant, 26 January 2026.

[36] Will Douglas Heaven, "OpenAI is throwing everything into building a fully automated researcher," *MIT Technology Review*, 20 March 2026, https://www.technologyreview.com/2026/03/20/1134438/openai-is-throwing-everything-into-building-a-fully-automated-researcher.

[37] Anthropic, "How AI is transforming work at Anthropic," 2 December 2025, https://www.anthropic.com/research/how-ai-is-transforming-work-at-anthropic.

at Google is AI-generated.[38] Some experts claim that the impact will be comparable to the effective 'workforces' of each frontier lab growing exponentially.[39]

2.3.1 Rationalising the pace of change

Questions remain regarding the extent to which widescale automation of AI R&D would directly affect the rate of AI progress and capability uplift.[40] For example, will AI R&D productivity gains outpace the increase in difficulty of making new discoveries? And will the need to test novel ideas in real-world experiments constrain the pace of change? Physical trials in domains like biology and chemistry take time, and AI scientists would need to integrate robotic systems to perform experiments in wet labs. On the other hand, if labs focus on AI R&D tasks where sample efficiency improves dramatically, and AI systems can master new skills based on very little data, we could see abrupt real-world impacts much quicker than anticipated.

Therefore, various scenarios are still possible, including oscillation between rapid progress and periods where unresolved bottlenecks cause a slowdown.[41] If there turn out to be no significant bottlenecks to the automation of AI R&D, governments' ability to prepare and react will be highly exposed.[42] These developments will largely take place behind closed doors[43] and are unlikely to happen in one discrete event, accentuating the element of strategic surprise further.[44]

But if relevant information was shared, there are elements that policymakers could potentially track: the fraction of frontier labs' compute budget dedicated to AI-directed experiments compared to human-directed ones;[45] and researchers' self-reported AI use

[38] Alphabet Investor Relations, "2025 Q4 Earnings Call," 4 February 2026, https://abc.xyz/investor/events/event-details/2026/2025-Q4-Earnings-Call-2026-Dr%5FC033hS6/default.aspx.
[39] Dean W. Ball, "On Recursive Self-Improvement (Part I)," *Hyperdimensional*, 5 February 2026, https://www.hyperdimensional.co/p/on-recursive-self-improvement-part.
[40] Helen Toner et al., *When AI Builds AI: Findings From a Workshop on Automation of AI R&D* (CSET Georgetown: January 2026), https://cset.georgetown.edu/publication/when-ai-builds-ai.
[41] Daniel Kokotajlo et al., *AI 2027* (AI Futures Project: April 2025), https://ai-2027.com/ai-2027.pdf; Arvind Narayanan and Sayash Kapoor, "AI as Normal Technology," *Knight First Amendment Institute at Columbia University*, 15 April 2025, https://knightcolumbia.org/content/ai-as-normal-technology; Sayash Kapoor et al., "Common Ground between AI 2027 & AI as Normal Technology," *Asterisk Magazine*, 12 November 2025, https://asteriskmag.substack.com/p/common-ground-between-ai-2027-and.
[42] Daniel Eth and Tom Davidson, *Will AI R&D Automation cause a Software Intelligence Explosion?* (Forethought: March 2025), https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.pdf.
[43] Charlotte Stix et al., *AI Behind Closed Doors: a Primer on The Governance of Internal Deployment* (Apollo Research: April 2025), https://arxiv.org/pdf/2504.12170.
[44] Ball (2026).
[45] Jean-Stanislas Denain and Cheryl Wu, "Final training runs account for a minority of R&D compute spending," *Epoch AI*, 23 March 2026, https://epoch.ai/gradient-updates/r-and-d-vs-training-compute.

patterns and productivity gains across different AI R&D task categories (12 potential metrics are presented in a paper by Alan Chan et al.).[46]

In some sense, the delegation of AI R&D to AI systems represents its own kind of paradigm shift.[47] Automated AI R&D raises questions over whether new oversight mechanisms are needed.[48] It is unclear whether such automation will likely favour developments within the existing paradigm, or whether fundamental changes to the nature of AI itself are likely to be unleashed.[49] Automated AI R&D would likely result in more rapid progress in certain research directions that lend themselves more easily to automation.[50]

For now, the generation of novel insights necessary to generate new hypotheses and set research agendas is still an area where frontier models fall short, and decisions about which research directions to prioritise remain a human endeavour.[51]

2.4 Challenges that need solving

Despite rapid developments in frontier AI, fundamental challenges remain. The AI field is replete with strong convictions about which of these challenges can be solved within the current paradigm, versus which require something fundamentally distinct – most survey respondents for this research fell in the latter category.

The constraints prioritised in this research build on the UK AI Security Institute's (AISI's) 2025 paper on 'Understanding AI Trajectories: Mapping the Limitations of Current AI Systems'.[52] These are summarised as follows:

Performance limitations on certain types of tasks:

- Performance on tasks that are hard to verify.
- Performance on long tasks.
- Performance in complex environments.

[46] Alan Chan et al., *Measuring AI R&D Automation* (GovAI; University of Oxford: March 2026), https://arxiv.org/pdf/2603.03992.
[47] Author interview with industry participant, 21 January 2026.
[48] Joe Benton et al., *Sabotage Evaluations for Frontier Models* (Anthropic: October 2024), https://arxiv.org/abs/2410.21514.
[49] Ball (2026).
[50] Author interview with government participant, 26 January 2026.
[51] Ball (2026).
[52] Max Heitmann et al., *Understanding AI Trajectories: Mapping the Limitations of Current AI Systems* (AI Security Institute: October 2025), https://www.aisi.gov.uk/research/understanding-ai-trajectories-mapping-the-limitations-of-current-ai-systems.

Reliability limitations:

- Sufficiently low error rate.
- Meta-awareness and calibration.

Adaptability limitations:

- Adaptability to the deployment environment.
- Continual learning and post-deployment.

Originality limitations:

- Original insights.

The next section exploring novel AI paradigms will assess the extent to which alternative approaches to AI architectures and models, learning and data, and hardware and compute address the challenges highlighted above.

3. Novel Paradigms: Pushing the Frontier?

This section covers research directions that could shift capability, cost or risk beyond the existing paradigm. The aim is to understand the supporting evidence behind these paradigms, their potential for direct impact on real systems, and the pre-conditions for their ability to scale.

This report organises novel paradigms into the following categories, while recognising that boundaries can blur (an agentic system, for example, may depend on retrieval scaffolding and deployment constraints as well as model design):

- **Architectures and models:** changes to model structure or system design that affect capability, reliability, or how models interact with tools and environments. Examples include agentic systems, world models, neuro-symbolic approaches, and structured retrieval such as GraphRAG.
- **Learning and data approaches:** changes to how models are trained, updated and grounded. This includes continual learning, embodied learning, and methods that improve deployment through better efficiency or distributed training setups.
- **Hardware and compute substrates:** changes to the compute stack that affect performance per watt, latency, and what can be deployed in constrained settings. This includes accelerators beyond GPUs, neuromorphic systems, and emerging approaches such as quantum machine learning and thermodynamic computing.

The table below provides an overview of the extent to which the 15 approaches assessed in this section address the eight research challenges highlighted at the end of Section 2.

Table 1. Indicative mapping from current limitations to novel paradigms and adjacent approaches

Paradigm/approach	Research challenges being addressed	Key takeaway
Agentic approaches and tool use	Performance on long tasks; performance in complex	Already commercially deployed as an extension of current frontier models.

	environments; some hard-to-verify tasks.	Increased error and risk surface.
Retrieval scaffolds including GraphRAG and structured retrieval	Performance on hard-to-verify tasks; lower error rates in bounded domains; meta-awareness and calibration through provenance.	Most useful in bounded domains, especially where current, traceable information matters.
World models and planning-centric autonomy	Performance on long tasks; performance in complex environments; adaptability to deployment environment.	Promising for robotics, simulation, and structured autonomy, but weaker in socially open-ended domains.
Neuro-symbolic AI and rule-guided inference	Performance on hard-to-verify tasks; sufficiently low error rate; meta-awareness and calibration in structured settings.	Most relevant where explicit rules, constraints, and auditability matter.
Diffusion and generative model variants	Selective performance gains (depending on the model family) in controllability, infilling, generation efficiency, and sequence handling.	Credible alternatives to autoregressive language modelling, with promise for controllability and efficiency; less evidence that they solve deployment reliability on their own.
State-space and hybrid post-transformer sequence models	Performance on long tasks; performance in complex sequential environments; some efficiency-related	Credible alternative or complement for long-context and sequential processing, with promise in efficiency and memory

	constraints around long-context processing.	use; limited evidence for broad replacement of transformers.
Recurrent, recursive and memory-augmented architectures	Performance on long tasks; some hard-to-verify tasks involving multi-step reasoning; state tracking and memory efficiency.	Potential for long-horizon reasoning and memory efficiency, but limited evidence for broader deployment.
Continual learning as a capability extender	Adaptability to deployment environment; continual learning and post-deployment updating.	Important for post-deployment adaptation, allowing models to update to new tasks, data and environments without full retraining.
Embodied AI and sim-to-real learning	Performance in complex environments; adaptability to deployment environment; some long-horizon decision-making in physical settings.	Important where grounded interaction matters, but reliable transfer from simulation to real environments and safety remain major constraints.
Lean and efficient models for constrained deployment	Mainly affects deployment under latency, power, privacy and cost constraints, though in specialised domains can match or outperform larger general-purpose models.	High near-term importance for adoption, specialisation and sovereign deployment; can be highly competitive in bounded domains even if not a general replacement for frontier models.

Distributed and federated architectures	Adaptability to deployment environment; continual learning and post-deployment updating under privacy and data-locality constraints.	More of a systems and governance response than a core capability shift.
Alternatives to GPUs in current compute stacks	Shapes what can scale by addressing energy-efficiency, latency, memory-movement, and cost bottlenecks.	Strategically important because it can lower the cost of training and deployment, but does not by itself solve the main model limitations.
Neuromorphic computing	Performance in edge environments; adaptability to deployment in low-power settings.	Specialised but credible for sensing, real-time control, and constrained edge deployment.
Quantum machine learning	No clear near-term mapping across the main current limitations.	Low readiness and limited operational evidence; one to be monitored.
Thermodynamic computing	No clear near-term mapping across the main current limitations, but potentially relevant to probabilistic and sampling-heavy AI workloads.	Early-stage hardware exists, with first chips taped out and a plausible near-term role as a specialised co-processor.

3.1 Architectures and models

3.1.1 Agentic approaches and tool use

Agentic systems embed a language model, or another controller, in a loop that can set goals, plan steps, call tools, observe outcomes, and revise behaviour over time. This gives frontier systems greater autonomy and ability to take actions, so has significant potential to impact capabilities and risks.

Frontier labs are investing heavily in this direction. OpenAI has released tool-using and computer-using agent systems such as Operator and the Responses API tool stack, Anthropic has deployed computer use for Claude, and Google DeepMind now frames Gemini as supporting agentic coding and tool use.

By linking models to tools, software and structured data sources, agentic systems can handle more complex multi-step tasks. They also change the failure surface: errors can propagate through tool calls, memory, permissions, and repeated steps.

The current evidence suggests both progress and brittleness. Benchmarks and demonstrations show strong improvements in task completion when planning, reflection, and tool use are added. At the same time, while the horizon over which agents can operate reliably is extending, errors can still compound over longer autonomous sequences, especially where the system loses track of the current task, optimises for the wrong short-term objective, or relies too heavily on brittle heuristics.[53]

[53] Xiao Liu et al., *AgentBench: Evaluating LLMs as Agents* (Tsinghua University; The Ohio State University; UC Berkeley: October 2025), https://arxiv.org/pdf/2308.03688; Guanzhi Wang et al., *Voyager: An Open-Ended Embodied Agent with Large Language Models* (NVIDIA; Caltech; UT Austin; Stanford; UW Madison: October 2023), https://arxiv.org/abs/2305.16291; Asaf Yehudai et al., *Survey on Evaluation of LLM-based Agents* (The Hebrew University of Jerusalem; IBM Research; Yale University: March 2025), https://arxiv.org/pdf/2503.16416.

Technological readiness

Tool-using systems are already in use, but the key question is how far they can remain predictable and governable as autonomy, task length, and workflow integration increase. This depends less on raw benchmark performance and more on the surrounding control and governance framework, including permissions, logging, evaluation harnesses, sandboxing, rollback procedures, runtime assurance, and clear lines of oversight and accountability. Future agentic deployment could therefore reshape security, governance and geopolitical risk even if progress in the underlying models only remains steady.[54]

3.1.2 Retrieval scaffolds including GraphRAG and structured retrieval

Retrieval-augmented systems improve model performance by connecting generation to external knowledge sources. Structured retrieval approaches, including GraphRAG, push this further by organising knowledge in forms that support multi-hop retrieval and reasoning, traceability and better provenance.

Frontier models already use retrieval-augmented setups widely in deployment, especially where answers need to be grounded in current or private data rather than model weights alone. Graph-based retrieval is more specialised, but it is part of the same broader shift towards external knowledge access and structured grounding for complex tasks.

Crucially, in policy, intelligence, science and enterprise settings, a system that retrieves the right information – and can show where it came from – may be more useful than a larger model with weaker real-world grounding.

Graph-based retrieval is particularly relevant where relationships matter. A graph can make it easier to connect entities, events, sources, and chains of evidence. This can improve

[54] Department for Science, Innovation and Technology, *Code of Practice for the Cyber Security of AI* (January 2025), https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai; AI Security Institute, *Frontier AI Trends Report* (2025); Darren Cofer et al., "Run-Time Assurance for Learning-Enabled Systems," in *NASA Formal Methods: 12th International Symposium* (Moffett Field: Association for Computing Machinery, 2020), 361-368, https://dl.acm.org/doi/10.1007/978-3-030-55754-6_21; Dung Phan et al., *A Component-Based Simplex Architecture for High-Assurance Cyber-Physical Systems* (April 2017), https://arxiv.org/pdf/1704.04759; J. Tanner Slagel et al., *A Formal Verification Framework for Runtime Assurance* (NASA Langley Research Center: 2024), https://shemesh.larc.nasa.gov/fm/papers/NFM2024-draft.pdf.

performance on questions that require linking multiple facts rather than retrieving one chunk of text, while also supporting a cleaner audit trail.[55]

Technological readiness

Structured retrieval should be treated as a capability enhancer. Performance depends on graph construction, curation quality, indexing choices, update frequency, provenance, and access controls. In bounded domains, it can materially improve reliability, and in some settings, it may support better data control by keeping sensitive information in external stores rather than in model weights (this benefit depends on the architecture and threat model). But a weak graph, outdated data, or poor provenance can produce misleading answers with a veneer of confidence, and retrieval also creates attack surfaces around prompt injection, exfiltration, access control, provenance, and the security of the external knowledge base itself.[56]

3.1.3 World models and planning-centric autonomy

World models seek to learn an internal representation of an environment that can support prediction, planning and control. In practical terms, they aim to help a system to answer questions such as: what state is the environment in, what is likely to happen next, and which action sequence is most likely to achieve the goal?[57]

World models therefore offer a potential route to more grounded understanding. They are still learned statistical models, but they are trained to represent environmental dynamics more explicitly, including how the environment changes over time in response to the system's actions. This is especially relevant for robotics, autonomous systems, simulation-heavy domains, and tasks where long-horizon planning matters. This direction is already visible in current work such as Meta's JEPA and V-JEPA line, which Yann LeCun has framed as part of a push towards systems that learn internal models of the world for understanding, prediction and planning, including robot control.[58]

There is also a live debate about whether explicit world-model architectures will be needed at all. Some researchers argue that something akin to world models may arise inside larger-

[55] Qinggang Zhang et al., *A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models* (September 2025), https://arxiv.org/abs/2501.13958; National Cyber Security Centre et al., *Guidelines for secure AI system development* (November 2023), https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development; Department for Science, Innovation and Technology (2025).
[56] Ibid.
[57] David Ha and Jürgen Schmidhuber, *World Models* (May 2018), https://arxiv.org/abs/1803.10122; Danijar Hafner et al., *Dream to Control: Learning Behaviors by Latent Imagination* (March 2020), https://arxiv.org/abs/1912.01603.
[58] Metz (2026).

scale systems as an emergent capability, rather than appearing through a distinct architectural shift.

The appeal of world models is partly about data efficiency. A system that can accurately simulate possible futures may need fewer real interactions to learn useful behaviour. These models also help connect learning to action: a model of dynamics can be used not just to predict what happens next, but to act in the world more effectively. This makes them important in the context of embodied systems and operations in structured environments.[59]

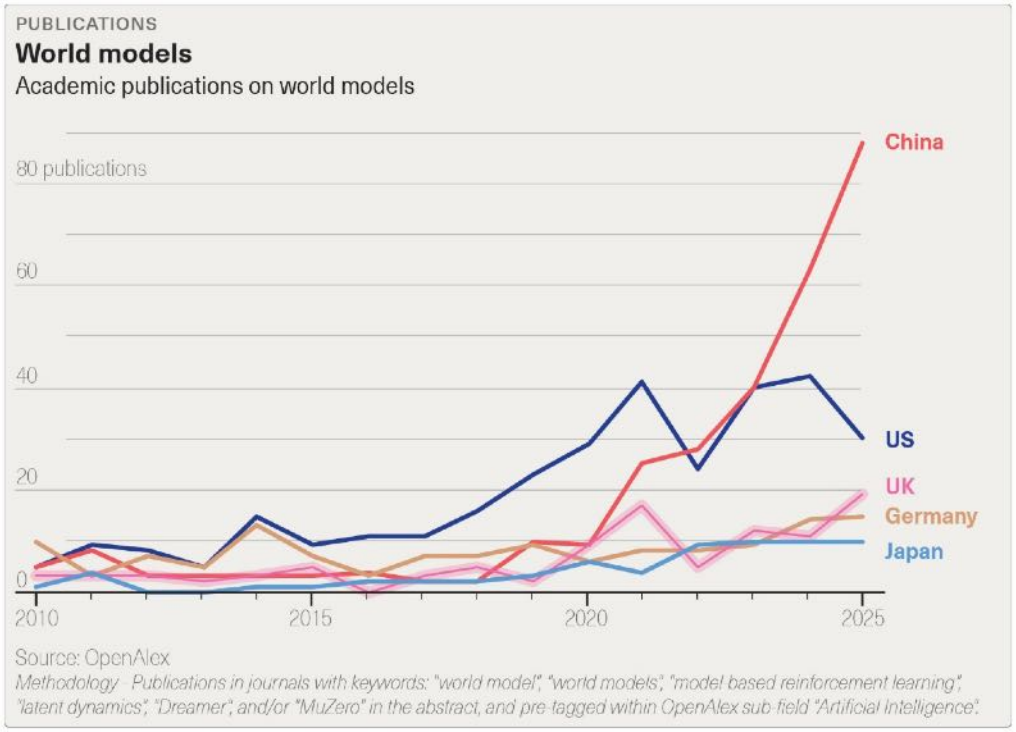
Technological readiness

The key challenge for world models is whether they can be shown to improve task performance in operationally relevant settings, and whether there is a credible path for monitoring and bounding error. Their main limitation is fidelity: a world model is only as useful as its representation of the environment and its ability to remain coherent over time. Model errors can accumulate, especially over longer horizons or under distribution shift, so a planning system built on a weak world model can become confidently wrong in ways that are hard to detect early. Their value is therefore likely to be highest in settings with relatively stable physical or statistical dynamics, including robotics, navigation, autonomy and other structured environments, rather than in tasks shaped heavily by human intent, politics, or strategic interaction. At present, world models look promising for specific settings and as part of a broader hybrid trajectory, especially when combined with embodied learning and simulation.[60]

[59] Danijar Hafner et al., *Mastering Diverse Domains through World Models* (Google DeepMind; University of Toronto: April 2024), https://arxiv.org/abs/2301.04104.

[60] Jingtao Ding et al., *Understanding World or Predicting Future? A Comprehensive Survey of World Models* (Tsinghua University: December 2025), https://arxiv.org/abs/2411.14499; Hafner et al. (2024).

Figure 4. Academic publications on world models since 2010



3.1.4 Neuro-symbolic AI and rule-guided inference

Neuro-symbolic AI[61] combines neural methods with symbolic structure such as rules, logic, constraints, graphs, or formal representations. The aim is to retain the pattern recognition strengths of machine learning while improving interpretability, controllability and reasoning over structured information.

Frontier labs are already using approaches that can plausibly be read this way. For example, Google DeepMind described AlphaGeometry 2 as a neuro-symbolic hybrid system built around a Gemini-based language model and a symbolic engine.[62]

The strongest case for neuro-symbolic methods is not as a replacement for large neural models, but as a practical approach for bounded settings where rules, constraints and auditability matter. Examples include structured reasoning over knowledge bases, logic-

[61] Some authors use the term narrowly for systems that tightly combine neural and symbolic components. Others use it more broadly for systems that pair learned models with rules, logic, or formal reasoning tools. In this report, we use the broader practical meaning.

[62] Yuri Chervonyi et al., *Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2* (Google DeepMind: December 2025), https://arxiv.org/abs/2502.03544.

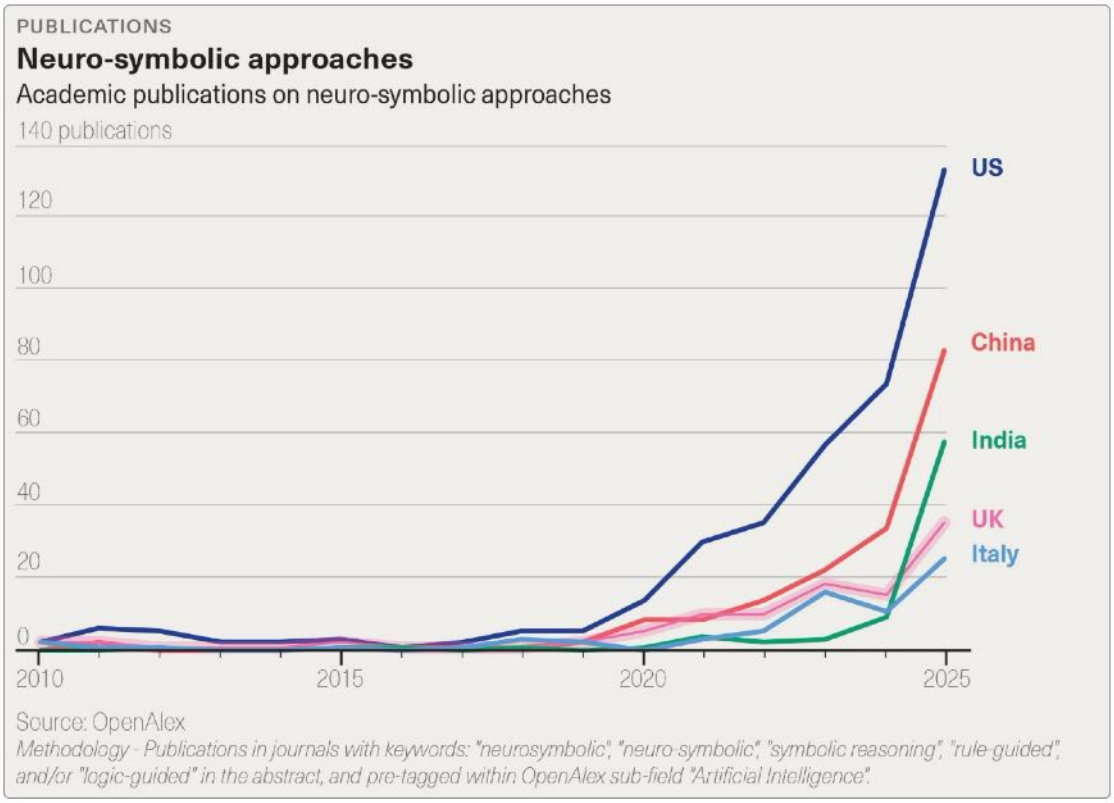
guided decision support, and domains such as formal verification and compliance-heavy decision support.[63]

Technological readiness

Neuro-symbolic AI is best viewed as an assurance-friendly path in selected domains rather than a general replacement for frontier models. Many neuro-symbolic systems work well in narrow or moderately structured settings but are harder to scale to open-ended, messy environments. There is also a risk that claims of interpretability are overstated if the symbolic layer is thin or only loosely connected to the main source of model behaviour.[64]

The neuro-symbolic publication trend in Figure 5 should be read carefully: it likely reflects broad academic and application-led interest across a wide field, rather than serving as a direct proxy for where frontier model labs are concentrating effort.

Figure 5. Academic publications on neuro-symbolic approaches since 2010



[63] Robin Manhaeve et al., "Neural probabilistic logic programming in DeepProbLog," *Artificial Intelligence* 298 (September 2021): 103504, https://doi.org/10.1016/j.artint.2021.103504.

[64] Artur d'Avila Garcez and Luís C. Lamb, "Neurosymbolic AI: the 3rd wave," *Artificial Intelligence Review* 56 (March 2023): 12387-12406, https://link.springer.com/article/10.1007/s10462-023-10448-w; Brandon C. Colelough and William Regli, *Neuro-Symbolic AI in 2024: A Systematic Review* (University of Maryland: April 2025), https://arxiv.org/abs/2501.05435.

3.1.5 Diffusion and generative model variants

Diffusion models are best known for image generation, but their significance is broader. In language, diffusion-based models replace token-by-token autoregressive generation (i.e. generating text word-by-word) with an iterative denoising process, and are becoming an increasingly credible alternative within generative and multimodal research. This category also covers model variants that may improve generation quality, controllability or efficiency in specific tasks.

Diffusion models demonstrate that model progress is not confined to one family. They have already reshaped image and media generation, are now appearing in language models such as Gemini Diffusion, and remain important across a wider set of scientific and technical applications.[65]

Technological readiness

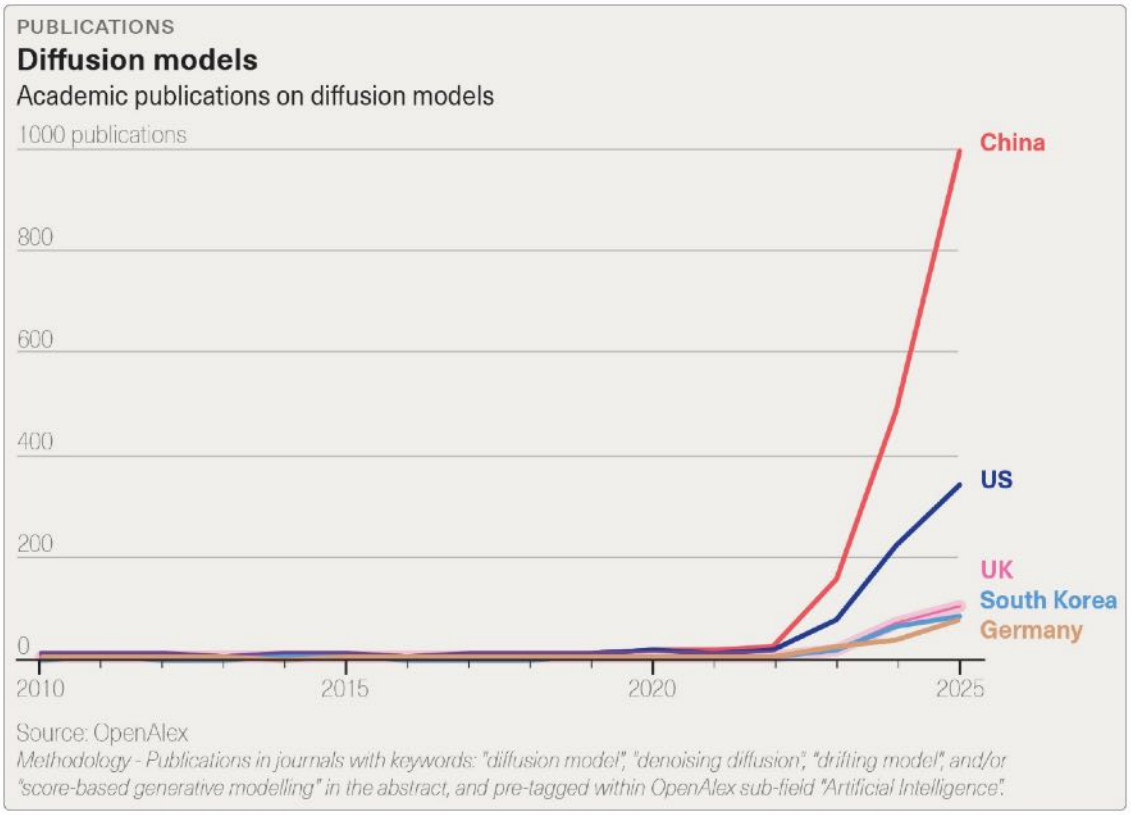
Diffusion should be treated as part of the broader picture of model diversification.[66] The main policy relevance lies in capability distribution and deployment patterns. A landscape made up of different specialised model families is harder to track and govern than one dominated by a single architecture. It also creates openings for countries or firms that are not at the frontier of one dominant approach to build strength in adjacent areas. While country-level publication counts show China leading in diffusion research (Figure 6), the frontier lab-level picture may look different: publication volume alone is not a reliable proxy for capability.[67]

[65] Google DeepMind, "Gemini Diffusion is our new experimental research model," 20 May 2025, https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-diffusion.

[66] Jonathan Ho, Ajay Jain and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851, https://dl.acm.org/doi/abs/10.5555/3495724.3496298; Robin Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models* (April 2022), https://arxiv.org/abs/2112.10752; Ling Yang et al., "Diffusion Models: A Comprehensive Survey of Methods and Applications," *ACM Computing Surveys* 56, no. 4 (November 2023): 1-39, https://dl.acm.org/doi/10.1145/3626235.

[67] Google DeepMind, "Gemini Diffusion," https://deepmind.google/models/gemini-diffusion.

Figure 6. Academic publications on diffusion models since 2010



3.1.6 State-space and hybrid post-transformer sequence models

State-space and hybrid post-transformer sequence models have attracted attention as alternatives or complements to standard transformer sequence modelling, especially for long-context processing and more efficient handling of sequential data. They aim to track and update internal state more efficiently than a standard transformer, and include pure or mostly state-space approaches such as Mamba, as well as hybrid architectures such as Jamba that combine transformer and state-space components.

Their importance lies in whether they can offer useful trade-offs in speed, memory, and context-handling relative to standard transformer-based models. If state-space methods can handle long-range dependencies more efficiently, they may become attractive in settings where cost, latency, or memory footprint are critical.

Technological readiness

At present, state-space models are an active and credible research direction, but not yet a replacement for transformers across general-purpose use cases. Their impact may emerge first in hybrid systems that combine state-space and transformer-style components – or in long-context and domain-specific sequence tasks – before any broader shift occurs. This would depend on credible software tooling, strong benchmarks, and robust deployment pathways.[68]

3.1.7 Recurrent, recursive and memory-augmented architectures

Recurrent, recursive and memory-augmented architectures remain worth monitoring because they target some of the main weaknesses of standard transformer stacks, especially state tracking, long-horizon reasoning, and memory efficiency. Recent examples suggest that these designs can be competitive on selected reasoning and efficiency benchmarks, but the evidence for broad deployment relevance is still limited. For now, they are better treated as an area to watch rather than as a clear alternative trajectory.[69]

3.2 Learning and data

3.2.1 Continual learning as a capability extender

Continual learning refers to updating a model over time without full retraining, while trying to preserve earlier capabilities and avoid losing skills or knowledge that the model had already learned. This is important because real-world systems rarely operate in a fixed environment: threats, tasks, interfaces and data distributions change, and systems that cannot adapt become stale quickly.

Continual learning sits close to the current frontier agenda. Many labs are interested in models that can incorporate new information and skills more fluidly. Ilya Sutskever has

[68] Albert Gu, Karan Goel and Christopher Ré, *Efficiently Modeling Long Sequences with Structured State Spaces* (Stanford University: August 2022), https://arxiv.org/abs/2111.00396; Albert Gu and Tri Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* (Carnegie Mellon University; Princeton University: May 2024), https://arxiv.org/abs/2312.00752; Simiao Zuo et al., *Efficient Long Sequence Modeling via State Space Augmented Transformer* (Georgia Institute of Technology; Microsoft: December 2022), https://arxiv.org/abs/2212.08136.

[69] Mostafa Dehghani et al., *Universal Transformers* (University of Amsterdam; DeepMind; Google Brain: March 2019), https://arxiv.org/abs/1807.03819; Bo Peng et al., *RWKV: Reinventing RNNs for the Transformer Era* (December 2023), https://arxiv.org/abs/2305.13048; Ali Behrouz, Peilin Zhong and Vahab Mirrokni, *Titans: Learning to Memorize at Test Time* (Google Research: December 2024), https://arxiv.org/abs/2501.00663.

argued that highly capable AI should be understood less as a finished system and more as a system that can *learn* new tasks quickly through deployment and experience.[70]

A system that can update incrementally could become more responsive, cheaper to maintain, and better suited to dynamic operational settings. It could also reduce dependence on long, expensive retraining cycles. For the UK policy context, this has implications for sovereignty and deployment, since the ability to adapt and update systems safely matters alongside continued access to leading foundation models.

However, incremental updates create new integrity and assurance questions: what exactly changed, what broke, and how can this be tested and rolled back? A model that learns continuously may also create a wider surface for poisoning, drift, and unintended performance losses in capabilities that previously worked well.

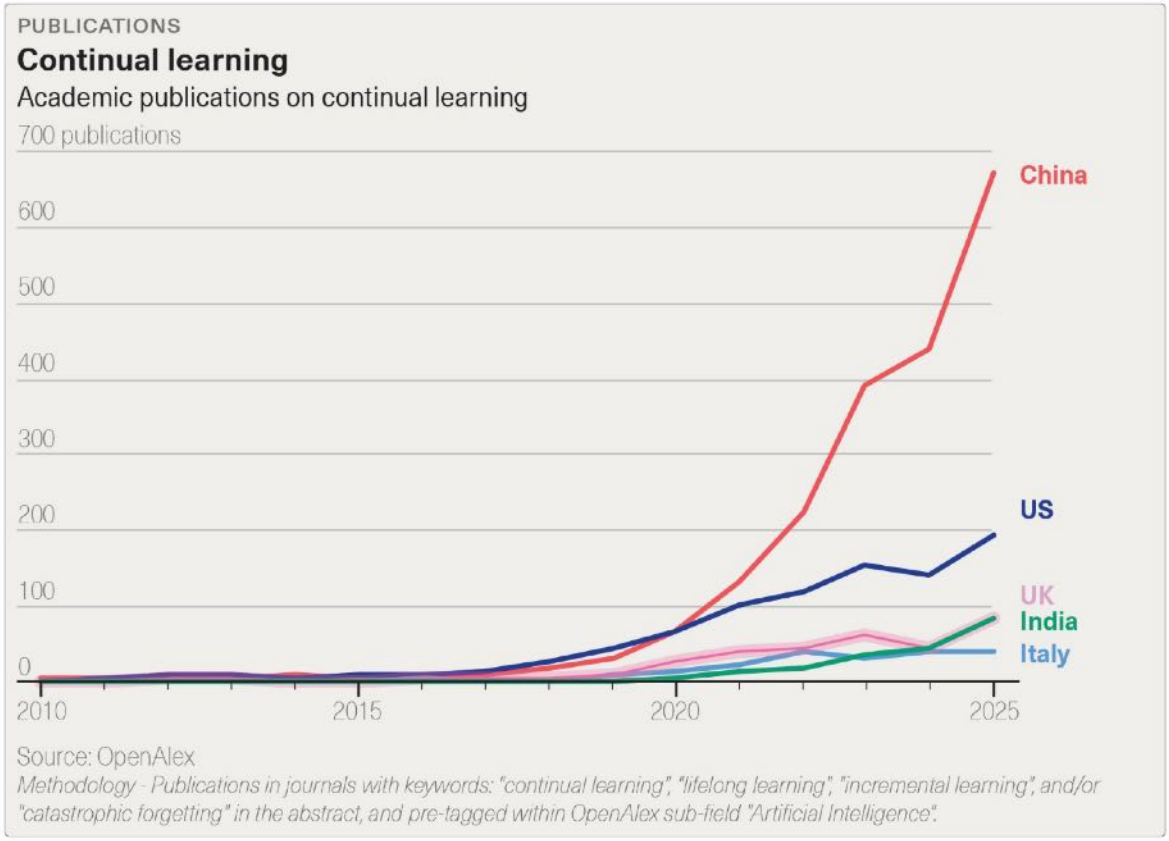
Technological readiness

Continual learning should be treated as strategically important but still constrained by evaluation and governance challenges. It is likely to matter more as an extension of the current paradigm than as a standalone alternative.[71]

[70] Patel (2025).

[71] Tongtong Wu et al., *Continual Learning for Large Language Models: A Survey* (Monash University; Griffith University: February 2024), https://arxiv.org/abs/2402.01364; Haizhou Shi et al., "Continual Learning of Large Language Models: A Comprehensive Survey," *ACM Computing Surveys* 58, no. 5 (November 2025): 1-42, https://doi.org/10.1145/3735633.

Figure 7. Academic publications on continual learning since 2010



3.2.2 Embodied AI and sim-to-real learning

Embodied AI focuses on learning through interaction with the physical world or a simulated environment, providing a route to more grounded intelligence. This aligns with Silver and Sutton's idea of an 'era of experience', in which capable AI systems learn increasingly through action and feedback rather than static human-generated data alone.[72]

Embodied learning offers one answer to concerns about the limits of text-based scaling, while linking AI progress more directly to robotics, autonomy, and operational systems by dealing with sensing, control, action and adaptation. It also connects closely to world models: a robot or embodied agent often needs an internal model of dynamics to predict what will happen next, test action sequences before acting, and reduce the amount of real-world trial and error needed for learning.

Simulation plays a central role here. In many settings, simulation is the only practical way to generate enough interaction data safely and cheaply. The question is whether capabilities

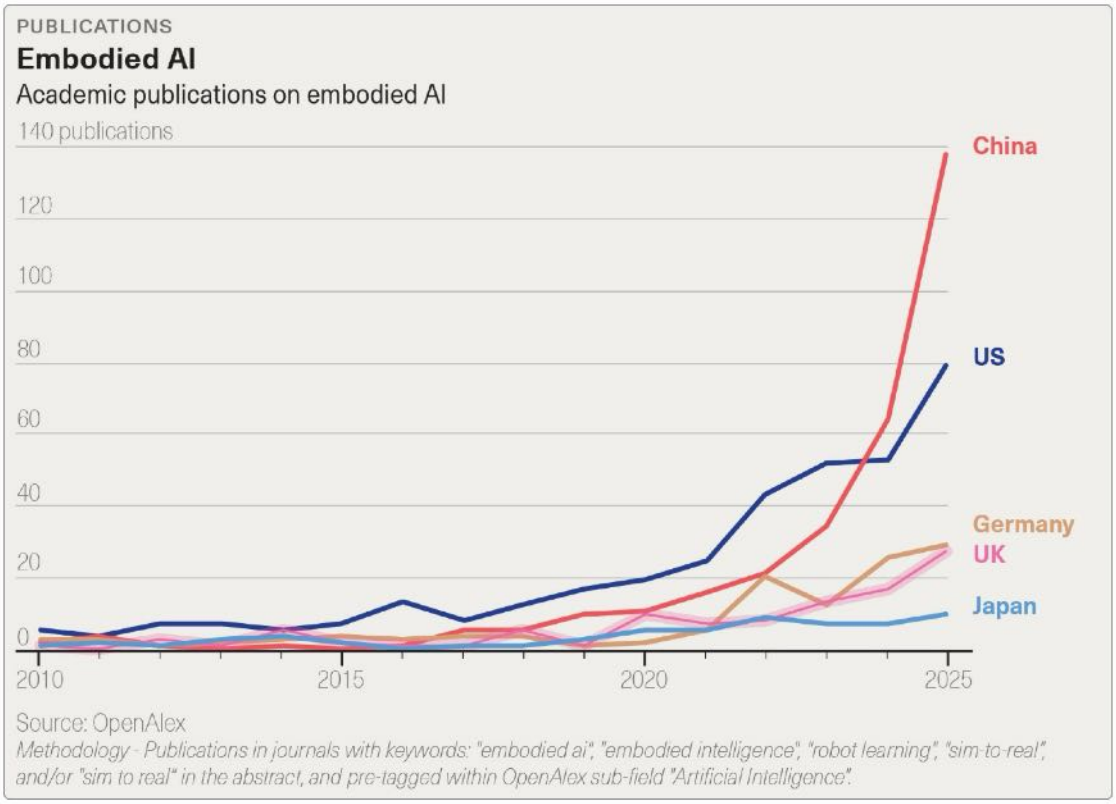
[72] David Silver and Richard Sutton, "Welcome to the Era of Experience," April 2025, https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf.

learned in simulation transfer reliably to real environments – this sim-to-real gap remains one of the central technical issues in the field.

Technological readiness

Embodied approaches are attractive because they can produce richer forms of adaptation and more grounded control. They also surface risks associated with physical systems: collisions, sensor occlusion or drift, actuator errors, unstable terrain or weather conditions, and real-world safety or mission consequences. Assurance is therefore critical for readiness.[73] Deployment at scale depends on reliable transfer, safety controls, and clear evaluation under realistic constraints.[74]

Figure 8. Academic publications on embodied AI since 2010



3.2.3 Lean and efficient models for constrained deployment

Lean and efficient models aim to reduce compute, memory, latency, or energy use, and include quantisation, pruning, distillation and other optimisation methods. Performance

[73] Cofer et al. (2020); Phan et al. (2017); Slagel et al. (2024).

[74] Zhiyuan Xu et al., *A Survey on Robotics with Foundation Models: toward Embodied AI* (Midea Group: February 2024), https://arxiv.org/abs/2402.02385; Ding et al. (2025).

alone is not enough to determine strategic value: a model that is slightly weaker on headline benchmarks may be far more useful if it can run cheaply, locally and reliably. This explains the relevance of lean models to edge systems, mobile devices, air-gapped settings, and sovereign deployment where cloud dependence is undesirable or impossible.

Efficiency methods can also shift the economics of adoption: reducing inference cost, allowing smaller organisations to deploy useful systems, and catalysing domains where latency, privacy or power constraints matter more than frontier-scale capability.

Technological readiness

Compression and optimisation of models can reduce broad generality, robustness, or observability in open-ended tasks. It can also complicate assurance if toolchains are immature or if the optimised model behaves differently in ways that standard evaluations do not fully capture. But that trade-off is not universal: in specialised or bounded domains, especially when combined with fine-tuning or retrieval, lean models can match and sometimes outperform larger general-purpose systems on targeted tasks. This is particularly relevant in agentic workflows, where latency, cost per call, and repeatability may matter more than broad general capability.

Lean and efficient models should be treated as a deployment multiplier. They may not be a novel paradigm in the strongest sense, but they can change which AI systems are viable in practice and which actors can field them.[75]

3.2.4 Distributed and federated architectures

Distributed and federated approaches change where learning happens and how models are updated or trained across devices, organisations, or compute clusters. At one end, this includes large-scale distributed training of frontier systems, where the training of a single model is spread over multiple sources of compute. At the other, it includes federated or split approaches designed to keep data local while still supporting collective learning.

Distributed and federated approaches are especially relevant to questions of sovereignty, privacy and deployment across public sector or critical infrastructure settings. They may

[75] Xubin Wang, Qing Li and Weijia Jia, *Cognitive Edge Computing: A Comprehensive Survey on Optimizing Large Models and AI Agents for Pervasive Deployment* (November 2025), https://arxiv.org/abs/2501.03265; Qualcomm, "Optimizing Your AI Model for the Edge," *Qualcomm Developer Blog*, 12 June 2025, https://www.qualcomm.com/developer/blog/2025/06/optimizing-your-ai-model-for-the-edge.

enable useful forms of AI adoption where raw data cannot be pooled centrally, or where operational systems are geographically and organisationally dispersed.

The challenge is that distributed learning is rarely free. Non-independent and non-identically distributed (non-IID) data; unreliable clients; communication and latency overheads; poisoning risks; and the complexity of audit and update control all make deployment harder. In large training runs, interconnect and synchronisation costs can become a real bottleneck.[76]

Technological readiness

Distributed and federated approaches should be treated as important enablers for selected settings rather than default solutions. Their value depends heavily on the deployment context, the security and governance constraints, and the quality of operational control around updates and monitoring.[77]

3.3 Hardware and compute

3.3.1 Alternatives to GPUs in current compute stacks

GPUs are not optimal across all the constraints defining the current AI trajectory, such as energy, memory movement, latency and cost. Some non-GPU accelerators are already commercially significant: Google's tensor processing unit (TPU) line is deeply embedded in large-scale AI deployment, and inference-optimised processors such as Groq's language processing unit (LPU) show that non-GPU accelerators are already carving out live deployment roles. Further out, more exploratory substrates, including photonic accelerators, wafer-scale systems, and in-memory computing approaches, are targeting specific bottlenecks but have not yet reached comparable deployment scale.[78]

---

[76] Samyam Rajbhandari et al., *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models* (Microsoft: May 2020), https://arxiv.org/abs/1910.02054; Mohammad Shoeybi et al., *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism* (Nvidia: March 2020), https://arxiv.org/abs/1909.08053.

[77] H. Brendan McMahan et al., *Communication-Efficient Learning of Deep Networks from Decentralized Data* (Google: January 2023), https://arxiv.org/abs/1602.05629; Peter Kairouz et al., *Advances and Open Problems in Federated Learning* (March 2021), https://arxiv.org/abs/1912.04977.

[78] Mark Lohmeyer, "Announcing the general availability of Trillium, our sixth-generation TPU", *Google Cloud Blog*, 11 December 2024, https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga; Google Cloud, "Cloud TPU release notes," *Cloud TPU documentation*, https://docs.cloud.google.com/tpu/docs/release-notes; Amin Vahdat, "Ironwood: The first Google TPU for the age of inference," *Google Blog*, 23 April 2025, https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/ironwood-tpu-age-of-inference; Groq, "What is a Language Processing Unit?," *Groq Blog*, 7 March 2025,

A hardware stack that can deliver better performance per watt or reduce system cost can materially change who is able to train and deploy advanced systems. It also affects supply chains and the extent to which countries or firms depend on a narrow set of vendors and manufacturing routes.

The strongest case for alternative substrates is in settings where the bottleneck is clear: in-memory compute targets data movement and memory bandwidth; photonic approaches target high-speed matrix operations; wafer-scale systems target throughput and interconnect efficiency for very large models. This is also why some alternatives have already gained real-world traction. Google's TPU programme is the clearest example: it shows that once a substrate is paired with a usable software stack and integrated into production workflows, it can move beyond a research curiosity and become part of a large-scale AI deployment strategy. The Trillium (v6e) TPU reached general availability in late 2024, and Google later introduced Ironwood as its seventh-generation TPU for large-scale inference.[79]

However, hardware gains only matter if they are matched by software tooling, compilers, debugging support, predictable integration, and an operational model that organisations can adopt. This helps explain why incumbent stacks retain their advantage: the Nvidia GPU-CUDA nexus combines mature hardware with a fast-moving software and engineering ecosystem, while alternative substrates must catch up on both the hardware and tooling side.

Technological readiness

Alternative compute approaches should be treated as strategically important, but uneven. Some are already shaping deployment (Google's TPU rollouts) while others remain plausible but still immature. The practical differentiator is whether the substrate has crossed from impressive demonstration to usable platform.[80]

---

https://groq.com/blog/the-groq-lpu-explained; Groq, "Groq and Nvidia Enter Non-Exclusive Inference Technology Licensing Agreement to Accelerate AI Inference at Global Scale," *Groq Newsroom*, 24 December 2025, https://groq.com/newsroom/groq-and-nvidia-enter-non-exclusive-inference-technology-licensing-agreement-to-accelerate-ai-inference-at-global-scale; Nvidia, "NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2026," *Nvidia Newsroom*, 25 February 2026, https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026.

[79] Lohmeyer (2024); Vahdat (2025).

[80] Norman P. Jouppi et al., *In-Datacenter Performance Analysis of a Tensor Processing Unit* (Google: April 2017), https://arxiv.org/abs/1704.04760; Ali Shafiee et al., *ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars* (University of Utah; Hewlett Packard Labs: October 2016), https://users.cs.utah.edu/~rajeev/pubs/isca16.pdf; Shiyue Hua et al., "An integrated large-scale photonic accelerator with

3.3.2 Neuromorphic computing

Neuromorphic computing draws on biological principles such as sparse, event-driven signalling and temporal processing, typically using specialised hardware designed for low-power, low-latency operation. It is best understood less as a direct rival to frontier AI systems than as a specialised pathway for improving efficiency in parts of the stack, especially inference at the edge.

Many strategically relevant applications are constrained by power, bandwidth and latency. Autonomous platforms, persistent sensing, and on-device perception cannot always rely on cloud access or large energy budgets. The evidence gathered for this project points to neuromorphic computing as most relevant in such settings.

A key point that emerged from research interviews is that neuromorphic computing is increasingly being shaped by the current AI paradigm. Its most credible near-term role is not to replace mainstream deep learning end-to-end, but to support more efficient execution of selected workloads, especially perception-heavy or 'always-on' tasks. Neuromorphic computing is less well matched to the dense, probabilistic and generative workloads that dominate the current frontier AI paradigm, which reinforces the view that its near-term role is likely to be specialised rather than general purpose.

However, neuromorphic systems still lack training methods with the maturity, flexibility and ecosystem support of standard backpropagation-based pipelines. In practice, this often means training conventional models first and then converting or adapting them for spiking or neuromorphic deployment.

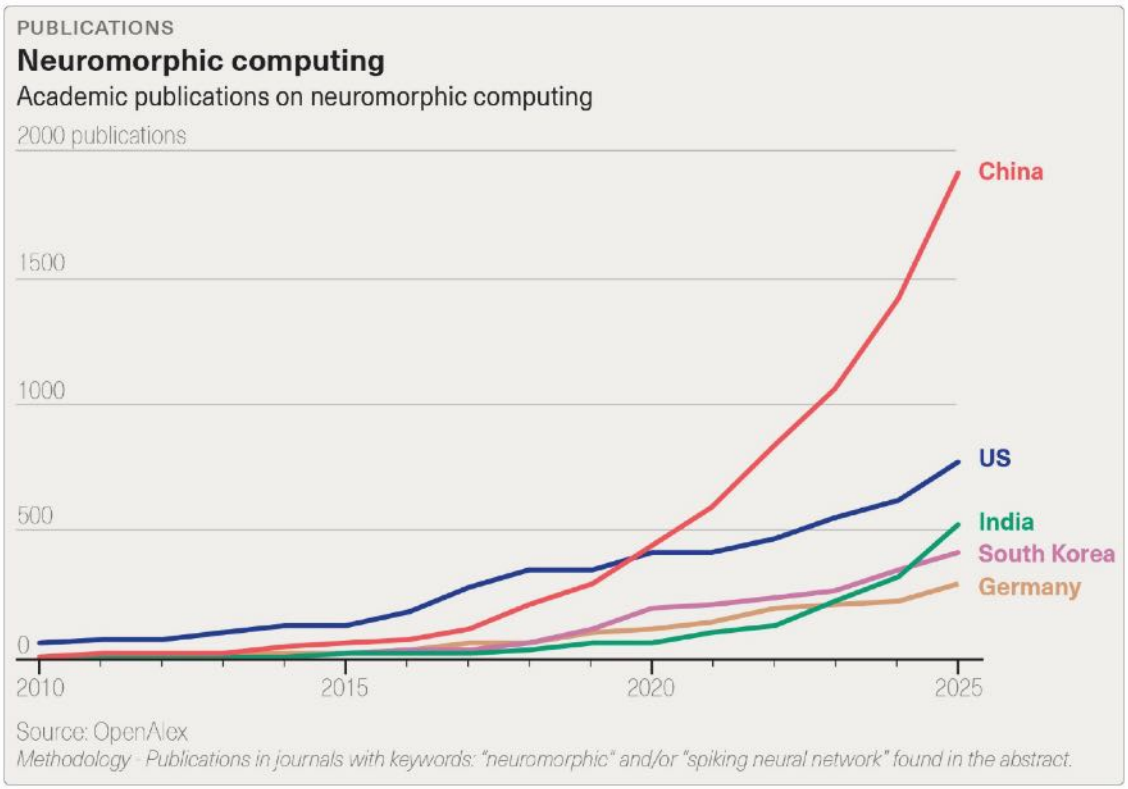
There are nevertheless concrete signs of ecosystem development. Research interviews pointed to platforms and firms such as Intel's Loihi, BrainChip, Innatera and SpiNNaker-related commercialisation efforts including SpiNNcloud as evidence that neuromorphic computing has moved beyond a purely academic niche. At the same time, these examples illustrate how deployment remains uneven, many systems are still at proof-of-concept or early deployment stage, and mainstream adoption remains far behind GPU-centred AI infrastructure.

ultralow latency," *Nature* 640 (April 2025): 361-367, https://www.nature.com/articles/s41586-025-08786-6; Congjie He et al., *WaferLLM: Large Language Model Inference at Wafer Scale* (University of Edinburgh; Microsoft Research: May 2025), https://arxiv.org/abs/2502.04563.

Technological readiness

Neuromorphic computing is best treated as a credible but bounded pathway. Its strongest near-term contribution is likely to come in constrained edge settings where sparse, event-driven processing matches the task and where energy efficiency is operationally decisive. The challenge is less whether neuromorphic ideas are promising in principle, and more whether they can be turned into usable, benchmarked and deployable systems.[81]

Figure 9. Academic publications on neuromorphic computing since 2010



3.3.3 Quantum machine learning

Quantum machine learning (QML) refers to approaches that use quantum circuits for learning, optimisation or simulation, often in hybrid quantum-classical workflows.

[81] Dhireesha Kudithipudi et al., "Neuromorphic computing at scale," *Nature* 637 (January 2025): 801-812, https://www.nature.com/articles/s41586-024-08253-8; Steve Furber, "Digital neuromorphic technology: current and future prospects," *National Science Review* 11, no. 5 (May 2024), https://academic.oup.com/nsr/article/11/5/nwad283/7342475; Steven Abreu et al., *Neuromorphic Principles for Efficient Large Language Models on Intel Loihi 2* (Intel Labs; UC Santa Cruz: March 2025), https://arxiv.org/abs/2503.18002; BrainChip, *Benchmarking AI Inference at the Edge* (January 2023), https://brainchip.com/wp-content/uploads/2023/01/BrainChip_Benchmarking-Edge-AI-Inference-1.pdf.

QML is often invoked in discussions regarding industrial strategy and long-range capability. It may also prove useful in narrow domains where optimisation or simulation structure matches what near-term quantum hardware can support.

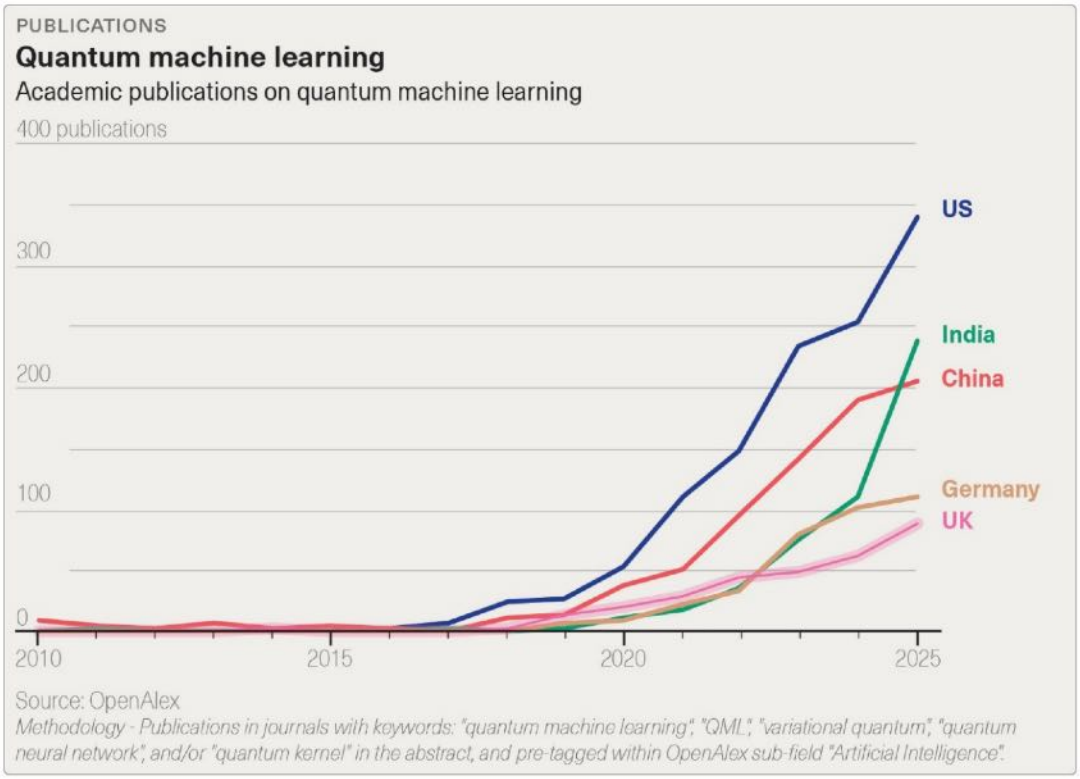
At present, the evidence for QML offering broad near-term advantage remains limited. Most work still sits closer to exploratory research than operational capability. Noise, limited qubit counts, data encoding overheads, and hardware constraints all narrow the range of realistic applications.

Technological readiness

Claims about QML sometimes run ahead of evidence, especially where comparisons are made against weak classical baselines or toy problems. QML cannot yet be treated as a major contributor to near-term AI capability unless much stronger applied evidence emerges.[82]

[82] Kamila Zaman et al., *A Survey on Quantum Machine Learning: Current Trends, Challenges, Opportunities, and the Road Ahead* (eBrain Lab: June 2025), https://arxiv.org/abs/2310.10315; Kavita R. Singh et al., "A Survey of Quantum Machine Learning: Understanding the Current Landscape and Future Opportunities," *Operations Research Forum* 6 (December 2025): 170, https://link.springer.com/article/10.1007/s43069-025-00569-z; Pradeep Lamichhane and Danda B. Rawat, "Quantum Machine Learning: Recent Advances, Challenges and Perspectives," *IEEE Access* 13 (2025), https://ieeexplore.ieee.org/abstract/document/11014055.

Figure 10. Academic publications on quantum machine learning since 2010



3.3.4 Thermodynamic computing

Thermodynamic computing uses noise and randomness as part of computation, rather than trying to eliminate them. This makes it potentially relevant to probabilistic and sampling-heavy AI workloads. Early hardware now exists, including the stochastic processing unit reported in *Nature Communications*, Normal Computing's CN101 tape-out, and Extropic's thermodynamic sampling units.[83]

The relevance of thermodynamic computing to this report is more forward-looking than immediate. Current prototypes demonstrate feasibility at small dimensions, but scaling to production-grade chips with high-speed interfaces and usable software integration remains a significant engineering task. It is also not yet proven that thermodynamic hardware can deliver consistent advantages over GPUs at the workload sizes that matter most for advanced AI systems.

[83] Denis Melanson et al., "Thermodynamic computing system for AI applications," *Nature Communications* 16 (April 2025): 3757, https://www.nature.com/articles/s41467-025-59011-x; Normal Computing, "Normal Computing Announces Tape-Out of World's First Thermodynamic Computing Chip," 12 August 2025, https://www.normalcomputing.com/blog/normal-computing-announces-tape-out-of-worlds-first-thermodynamic-computing-chip; Extropic, "Thermodynamic Hardware," https://extropic.ai/hardware.

Technological readiness

Thermodynamic computing should be treated as an early but credible emerging pathway. Its most plausible near-term role is as a specialised co-processor for probabilistic and sampling-heavy workloads rather than as a general-purpose replacement for GPUs. Reversible computing is a related but distinct and more speculative domain, focused on reducing dissipation by preserving information through computation.[84]

3.4 Summary and implications

Taken together, the paradigms and adjacent approaches in this section do not point to one clear successor to the current frontier approach. This was echoed by survey data, which found a wide range of views on the likelihood of capability impact by 2030 across these paradigms. Regarding the question of what would shift their thinking about a paradigm's relevance, many responses clustered around two key factors: superior performance on out-of-distribution reasoning or causal interventions, and evidence of major gains in sample efficiency.

The stronger pattern across this section is diversification. **Progress is likely to come from a mixture of extensions, hybrids and specialised pathways rather than from a single clear replacement**. This suggests that the future AI landscape may become more varied in architecture, training method, deployment model, and hardware base.

Technological readiness should not be judged only by performance on headline frontier benchmarks. Several of the most important shifts discussed relate to deployment, integration and assurance rather than pure model capability. Agentic systems matter because they connect models to tools and workflows, but also widen the surface for error and misuse. Lean models matter because they make deployment feasible in constrained settings. Continually updated systems matter because they create new integrity, rollback and monitoring questions. Neuromorphic and alternative hardware matter because they can shift the cost and location of computation, but only if they are supported by usable tooling and integration.

[84] Rolf Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development* 5, no. 3 (July 1961): 183-191, https://ieeexplore.ieee.org/document/5392446; Charles H. Bennett, "Logical Reversibility of Computation," *IBM Journal of Research and Development* 17, no. 6 (November 1973): 525-532, https://dl.acm.org/doi/10.1147/rd.176.0525; Melanson et al. (2025); Normal Computing (2025); Extropic, "Thermodynamic Computing: From Zero to One," 29 October 2025, https://extropic.ai/writing/thermodynamic-computing-from-zero-to-one.

Boundaries between paradigms are likely to remain blurred. Continual learning is best understood as a capability extender to the current paradigm. World models may combine with embodied systems. Structured retrieval can sit inside agentic systems, and agentic systems can also incorporate world models, memory, and retrieval layers. Neuro-symbolic methods may serve as a control layer rather than a full architecture shift. These overlaps are why this report frames the analysis in terms of understanding what each approach changes, what it depends on, and where it becomes strategically meaningful.

The unit of assurance is shifting from the model to the full system. Once a model is embedded in tools, retrieval, memory, external data sources, or continual update loops, failure can arise from orchestration, permissions, provenance, rollback and monitoring rather than from the base model alone. A paradigm may therefore look promising in isolation but still prove operationally brittle in deployment. The evaluation question is no longer only whether the model performs well, but whether the wider stack can be tested, monitored, governed and kept within acceptable bounds in real-world use.

For the UK, the strategic lesson is that **preparedness should be built for a range of futures rather than for one assumed winner**. The current UK Government position of diversification through a larger number of small bets (for example through the Sovereign AI Fund[85] and the new Fundamental AI Research Lab[86]) and maximising flexibility under different futures is characteristic of a UK-sized economy with limited fiscal headroom. **This places emphasis on identifying where the UK can add value outside the narrow race for the largest foundation model, for example by building capacity around deployment, assurance and systems integration**.

[85] "UK Sovereign AI Fund," https://sovereignai.gov.uk.

[86] Department for Science, Innovation and Technology, UK Research and Innovation and Kanishka Narayan MP, "Government to create new lab to keep UK in the fast lane on AI breakthroughs," 4 March 2026, https://www.gov.uk/government/news/government-to-create-new-lab-to-keep-uk-in-the-fast-lane-on-ai-breakthroughs.

4. Navigating Uncertain AI Trajectories

The following section explores the security and strategic implications of uncertain AI trajectories. It begins by outlining the US and China's global leadership and the 'moat' established by frontier companies, before considering how governments should understand technological asymmetries, manage deployment risks, and build capabilities for long-term preparedness. It finishes by focusing on what the UK should do in the face of uncertainty, positing measures across four core policy priority areas.

4.1 The global picture and frontier moat

Throughout this report, various data points have emphasised US leadership at the AI frontier, with China close behind. These two countries employ 70% of the world's top machine learning researchers[87] and command 90% of AI training compute.[88] US companies, national labs and academia have traditionally been the home of big research bets, including in domains like LLMs, diffusion models, optical computing and neuromorphic computing.

Meanwhile, China has excelled in building on this research, seen through its innovations in post-transformer architectures and hybrid transformers. China's success is also characterised by its ability to cross the 'valley of death',[89] pulling research capability through into leadership in critical manufacturing sectors such as batteries, solar panels and electric vehicles.[90] China's strengths as a 'fast follower' can be seen in Section 3 (Figures 4 and 6-9), where publications from China overtake those from the US following an initial lag.

It has been argued that resource constraints driven by US export controls have played an important role in stimulating Chinese innovation. DeepSeek's R1 and V3 models are the oft-cited examples in this regard, although the efficiency gains that DeepSeek achieved may also be reflective of broader algorithmic trends, and export controls have still made aspects of Chinese research and deployment more difficult.[91]

[87] It is likely that this statistic counts by employer location rather than nationality, which would include non-US/Chinese nationals working in the two countries.

[88] Sam Winter-Levy and Anton Leicht, "The AI Divide," *Foreign Affairs*, 10 February 2026, https://www.foreignaffairs.com/united-states/ai-divide.

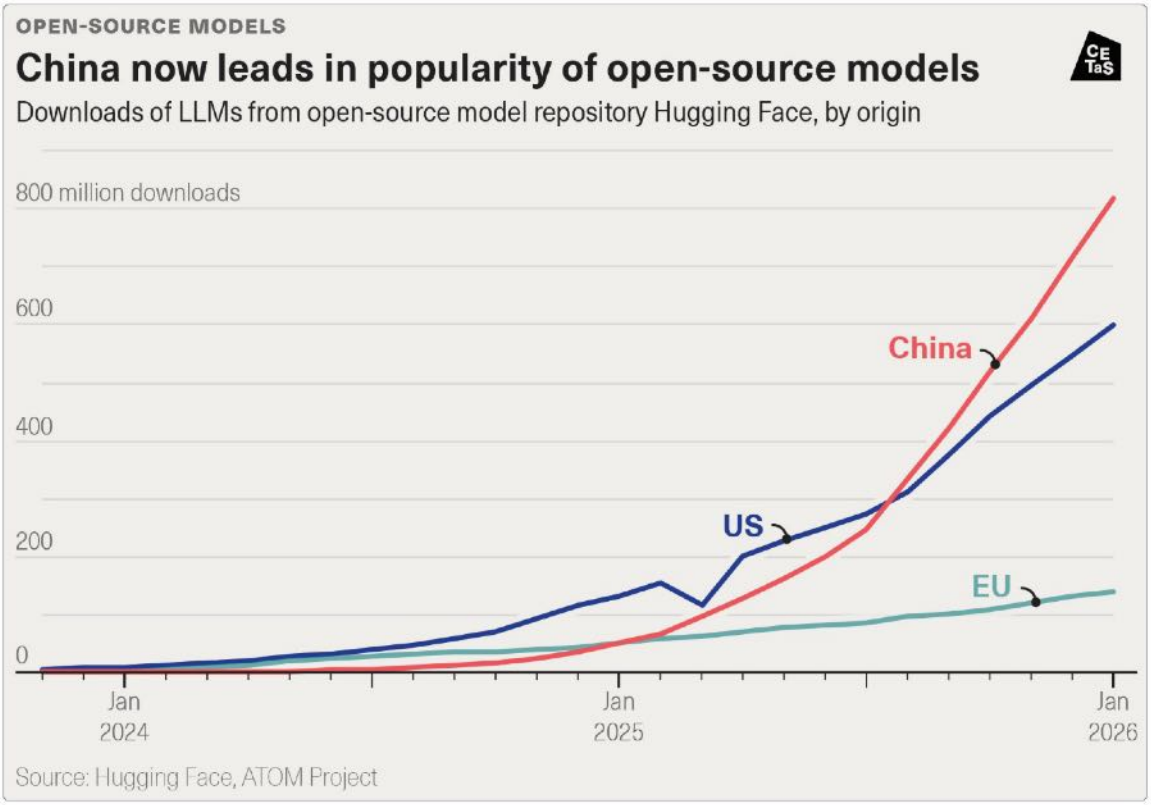
[89] Yue Yin, Ming Yan and Qiushi Zhan, "Crossing the valley of death: Network structure, government subsidies and innovation diffusion of industrial clusters," *Technology in Society* 71 (November 2022), https://doi.org/10.1016/j.techsoc.2022.102119.

[90] Author interview with academic participant, 15 January 2026; Laura Bicker, "As Trump retreats from climate goals, China is becoming a green superpower," *BBC News*, 18 February 2026, https://www.bbc.co.uk/news/resources/idt-8d2b6944-4f7a-45b4-96fd-2d92499ff97d.

[91] Author interview with government participant, 11 February 2026.

Additionally, China's play for the open-source ecosystem has been central to its strategic approach:[92] Chinese open models are now downloaded more than their US counterparts (Figure 11).[93]

Figure 11. Downloads of open-source large language models



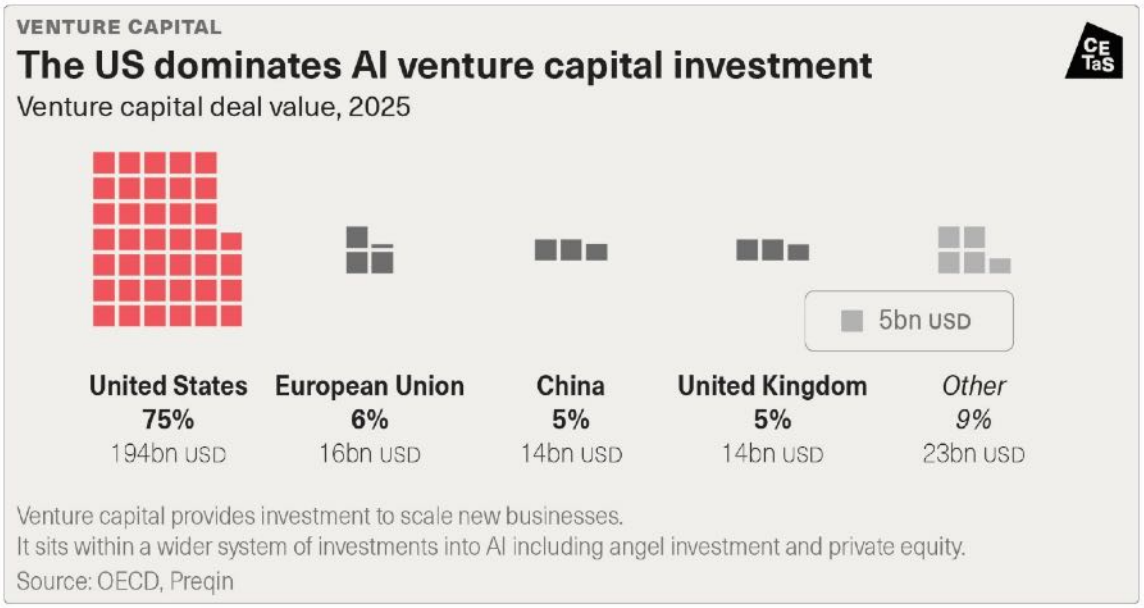
Yet China's leadership of the open-source ecosystem still trails the frontier led by US companies by several months. As well as pursuing the performance dividend from continued scaling, the US is well positioned to develop complementary approaches in algorithmic, architectural and learning paradigms.[94] For example, Google has long maintained a quantum research portfolio,[95] while Elon Musk's integration of SpaceX with xAI has been linked with grand ambitions of data centres in space.[96]

[92] Nathan Lambert, "Towards American Truly Open Models: The ATOM Project," *Interconnects*, 4 August 2025, https://www.interconnects.ai/p/atom-project.
[93] Nathan Lambert, "8 plots that explain the state of open models," *Interconnects*, 7 January 2026, https://www.interconnects.ai/p/8-plots-that-explain-the-state-of.
[94] Author interview with industry participant, 20 January 2026.
[95] Google Research, "Quantum Computing," *Research areas*, https://research.google/research-areas/quantum-computing.
[96] Akash Sriram and Joey Roulette, "Musk's mega-merger of SpaceX and xAI bets on sci-fi future of data centers in space," *Reuters*, 4 February 2026, https://www.reuters.com/business/aerospace-defense/musks-mega-merger-spacex-xai-bets-sci-fi-future-data-centers-space-2026-02-04.

The most innovative new paradigms may yet emerge within US labs, and even if not, the difficulty of scaling commercially in places like Europe means that (without state intervention) promising startups would face strong pressure towards foreign acquisition.[97]

In 2025, the US saw an order of magnitude more AI venture capital investment than European peers (Figure 12), with proportionally more in later-stage ventures.[98]

Figure 12. Venture capital investment in AI



US labs have a strong talent base and significant amounts of infrastructure about to come online in the next couple of years,[99] so the possibility of new hardware breakthroughs further entrenching their lead is highly plausible.

4.2 Risks from new paradigms

4.2.1 Technological asymmetries as a destabiliser

Asymmetrical AI capabilities between adversaries resulting from more globally diverse technical approaches could exacerbate deterrence and competition dynamics in military,

[97] Author interview with industry participant, 20 January 2026.
[98] OECD, *Venture capital investments in artificial intelligence through 2025* (February 2026), https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/venture-capital-investments-in-artificial-intelligence-through-2025_3bcb227f/a13752f5-en.pdf.
[99] Katie Tarasov, "Nvidia's new AI system Vera Rubin is 10 times more efficient than its predecessor – here's a first look," *CNBC*, 25 February 2026, https://www.cnbc.com/2026/02/25/first-look-at-nvidias-ai-system-vera-rubin-and-how-it-beats-blackwell.html.

economic and geopolitical contexts. Crucially, asymmetries could threaten to lower the threshold for conflict among military powers.

One example given by an interviewee focused on world models: if these evolve non-uniformly, there will be a disparity in capabilities of autonomous systems such as drones, and this advantage would be reinforced further among actors that develop more energy-efficient forms of AI, enabling sensor processing to be done at lower power levels.[100]

Such asymmetries emphasise the risk of relying on sub-state-of-the-art capabilities for national security purposes.[101] Governments' threat models must assume that some malicious actors will be operating the most capable forms of AI, and these actors may be unconstrained by political factors such as supporting homegrown AI companies.[102]

One interviewee outlined the following five questions as a way of understanding how such asymmetries may arise through the diffusion of novel paradigms.[103] For most paradigms, the answers to these questions are likely to exist on a spectrum.

- Is a new paradigm making it easier to be a fast follower?
- Is a new paradigm making models more commoditised or not?
- Is a new paradigm enabling step changes in capability?
- Is a new paradigm changing the barrier to entry (to being a frontier lab)?
- Is a new paradigm enabling a more permissive policy environment?

4.2.2 Managing deployment risks

Governments will also want to understand the risks from systems deployed in the most high-stakes, high-trust environments. There is a tension: on one hand, it may be easier to reason about the security properties (including risks such as data exfiltration and poisoning) for simpler or more constrained systems.[104] On the other hand, lower-capability models can also introduce different operational risks, particularly if adversaries are continually operating with the most powerful capabilities. Resolving this tension will involve identifying tasks where specialised models are more appropriate, and understanding how fine-tuning and

[100] Author interview with academic participant, 15 January 2026.
[101] Anton Leicht, "Import Imperatives," *Threading the Needle*, 6 February 2026, https://writing.antonleicht.me/p/import-imperatives.
[102] Author interview with academic participant, 5 February 2026.
[103] Author interview with government participant, 9 February 2026.
[104] Author interview with academic participant, 28 January 2026.

domain-specific post-training can ensure that general models are still applicable to sensitive contexts such as industrial control systems.

The arrival of the agentic paradigm has signalled alarm bells from an AI security perspective. Simon Willison coined the 'lethal trifecta for AI agents', referring to: i) providing an agent with access to private data, ii) exposing it to untrusted content, and iii) allowing it the ability to take actions in the world.[105] Agents combining these features make it much easier for an attacker to steal data. As a result, the more agentic that workflows become (an AI system gaining the ability to transfer money, for example), the stronger the incentives for attackers.[106] The pursuit for performance and edge in a contested world means that such concerns are less likely to be prioritised. Indeed, most survey respondents voted for 'tool-use and agentic autonomy misuse' as a priority 'new or worsened attack surface'.

The rollout of agentic systems in the national security context must be cognisant of two critical challenges: i) logical integrity, accountability and auditability, and ii) out-of-distribution reliability (ability to generalise).

Table 2. Key hurdles to securing agentic deployment in the existing paradigm

Technical challenges	**Deployment vulnerabilities**	**Current approaches to solving challenge**
Logical integrity, accountability and auditability	• **Automation bias** and the **consistency of high-stakes decision-making**.[107] • Additional risk of compounding of errors/hallucinations in an agentic context.[108] • Current assurance and oversight not being well suited to agentic systems.	• **Deliberative reasoning processes.** • **Process-based reward models** that train verifiers to grade logical steps of a model rather than the final answer. • **Post-hoc saliency mapping** that

[105] Simon Willison, "The lethal trifecta for AI agents: private data, untrusted content, and external communication," *Simon Willison's Weblog*, 16 June 2025, https://simonwillison.net/2025/Jun/16/the-lethal-trifecta.
[106] CETaS workshop, 29 October 2025.
[107] Author interview with government participant, 11 February 2026.
[108] CETaS workshop, 29 October 2025.

		visualises which data points most influenced a model's decision.
Out-of-distribution reliability (ability to generalise)	• **Verifiability beyond narrowly verifiable tasks and overcoming the friction of the real world**.[109] Models fail when a real-world situation does not match their training data (e.g. a novel cyber threat or financial anomaly). This is particularly salient for agentic models acting in a large action space.	• **Leveraging RL agents** to find model vulnerabilities and turning these failures into synthetic training examples. • Greater emphasis on **inference-time deliberation** to test assumptions against logical rules.

It would be prudent for work on AI security and risks to not just focus on risk from today's systems, but look at the potential risks from emerging approaches. This will require additional resource, and a willingness to place bets: some of these risks will not be realised.

4.3 UK policy priorities

There are two fundamental aspects to UK AI competitiveness. The first is building the domestic capabilities needed to absorb and apply AI breakthroughs, which means strengthening anticipatory capability, skills, infrastructure, and adoption at scale. The second is maintaining access to frontier capabilities where the UK remains dependent on foreign providers.

Access to frontier AI systems depends on infrastructure controlled by a small number of non-UK firms. In a more turbulent geopolitical environment, access conditions could tighten, especially for high-sensitivity national security use cases. UK policy therefore must do two things simultaneously: preserve access to the best available capabilities, while also building the domestic stack needed to adapt models to high-priority contexts.[110] These measures will

[109] Author interview with academic participant, 5 February 2026; Author interview with government participant, 9 February 2026.
[110] Winter-Levy and Leicht (2026).

enable the UK to become more responsive when AI breakthroughs occur, rather than attempting to predict the next breakthrough.

4.3.1 Enhancing anticipatory capabilities

Policymakers must design systems that can cope with surprise, rather than systems that aim to engineer this out. Much planning and investment has gone into building up *anticipatory capabilities* within the UK Government, with increasingly sophisticated approaches to picking up signals of technological progress.[111]

AISI is well positioned to identify near-term tactical risks, while the responsibility for highlighting strategic risks is shared with analysis and assessment bodies in Whitehall. For example, the Government Office for Science leads on technology forecasting work by cohering a cross-government emerging technology community. It is imperative that there is sufficient capacity to focus on developments beyond the current paradigm in an iterative, ongoing manner.

In intelligence assessment, analysis is often focused on discrete projects, and expertise is primarily cohered through roundtable consultation. Systematically tracking developments requires infrastructure and expertise in data science and engineering to make more use of available data, and making sense of downstream impacts also requires expertise in social science disciplines. Having these foundations in place will enable better integration of advanced horizon-scanning techniques that aggregate insight from large bodies of disparate information, such as crowd-forecasting, surveys, publication data, commercial data and market analysis.[112]

Much crucial information concerning the current paradigm and its future direction is located within AI labs. This is why establishing trusted information flows is important. From the labs, governments want to know about capabilities, deployment risks and incidents. On the other hand, labs would benefit from a more structured understanding of the threat landscape and ways to protect their research.

This report proposes the following recommendations for UK Government to enhance anticipatory capability:

1) Task existing horizon-scanning and technical assessment functions to **monitor the 15 novel paradigms** identified in this report **against the eight key technical**

[111] Author interview with government participant, 19 February 2026.
[112] Author interview with government participant, 11 February 2026.

challenges, producing **regular updates** rather than one-off reviews. These efforts should leverage data science to draw on a wider range of information sources, to complement expert opinion. The list of novel paradigms should be periodically reviewed.

2) Establish a standing **cross-disciplinary national security AI assessment capability** that brings together expertise from a broad range of **AI subfields (including alternative paradigms), hardware fields, biotechnology, cognitive science, materials science and physics, geoeconomics, and supply chains**, so that horizon-scanning and technical assessment are not confined to general-purpose model progress alone.

3) Establish **secure channels for sharing threat data, evaluation findings, and cybersecurity guidance** across government, frontier labs, and trusted partners, supported by secure testing environments for sensitive AI systems.

4.3.2 Building a deep skills base

Regardless of the dominant paradigm, having a deep skills base across software and hardware will ensure that there is a supply chain of talent that can realise the visions being funded by research bodies.[113]

The UK does not have many experts with direct experience of developing and maintaining large-scale AI models. UKRI Centres for Doctoral Training (CDTs) produce a sizeable number of doctoral researchers in AI, but those programmes need better access to opportunities to train, adapt and evaluate medium-to-large-scale models.

From a hardware perspective, many of the alternative compute paradigms assessed in this report require niche skills compared to the software space.[114] Some companies have recognised the importance of the software-hardware interface for future paradigms: Amazon and IBM have well-defined APIs in their quantum research, so that when the technology is mature enough, people will have the right skills to program to it.[115] The importance of this synchronised approach was reinforced in conversations with

[113] Author interview with academic participant, 15 January 2026.
[114] Ibid.
[115] Jordan Sullivan, "Introducing the Qiskit provider for Amazon Braket," *AWS Quantum Technologies Blog*, 17 June 2022, https://aws.amazon.com/blogs/quantum-computing/introducing-the-qiskit-provider-for-amazon-braket.

neuromorphic computing experts, one of whom pointed to an imbalance between researchers in the UK working at the device/material level versus the systems level.[116]

This report proposes the following recommendations for building a deeper UK skills base:

1) **Pre-training skills**: create a sovereign AI training pathway that gives selected UK researchers and engineers **hands-on access to medium-sized base models**, **compute**, and **evaluation environments** to boost experience of building, maintaining and using large-scale generative models. There should be a particular focus on developing skills in pre-training.

2) **Transferable skills**: to increase readiness to pivot, the UK should invest in transferable technical skills throughout the AI ecosystem. This includes within technical and policy areas of government, through the academic pipeline (from schools through to PhDs and beyond), and through support to industry.

3) **Specialist skills**: build targeted capability in niche subfields where future paradigms depend on scarce systems expertise, especially at the software-hardware interface. The Advanced Research and Invention Agency (ARIA), UKRI, and sovereign AI programmes should back these areas, where the UK risks missing critical capability bottlenecks.

4.3.3 Investing in supporting infrastructure

Without the appropriate investment in infrastructure and scale, developing new paradigms will ultimately be in vain.[117] Particular attention should be paid to infrastructure that would allow the UK to pivot in response to breakthroughs.

Although frontier pre-training places by far the heaviest demands on power infrastructure, large-scale inference and fine-tuning still require reliable, affordable energy and grid capacity.

While developing a full frontier-scale stack domestically may not be feasible, an inference-only stack would leave the UK highly dependent on foreign providers. A *fine-tuning stack* offers the most plausible middle path: it would allow adaptation of non-UK models to domain-specific data to build specialised tools. Complementing this approach, a focus on

[116] Author interview with academic participant, 19 January 2026.
[117] Author interview with academic participant, 2 February 2026.

domestic lean-AI and the development of a domestic medium-sized base model should increase the opportunities to deploy UK models.

Datasets are another aspect of critical AI infrastructure. The UK's strengths in financial and professional services position it well to capitalise on data commoditisation and its emerging role as a strategic asset within the AI ecosystem.[118] The UK has largely been a producer rather than a consumer of its own data: existing mechanisms to share and license data to UK AI developers are inadequate. China, for example, has succeeded in creating an exchange for trading large data assets.[119]

A federated set of simulators and reference environments could also be developed to train, test and evaluate agent-based AI systems for high-value use cases. These could include realistic reproductions of critical national infrastructure (CNI) or manufacturing systems, enabling novel approaches to cyber resilience to be tested at scale before deployment.

This report proposes the following recommendations for investing in supporting infrastructure:

1) Build a **fine-tuning AI stack** that allows adaptation to domain-specific data to build **specialised tools.**

2) Develop a UK **medium-sized base model** and the compute, tooling and data needed for top UK teams to experiment and build on it.

3) Ensure that UK compute strategy accounts for the **diversification of AI hardware beyond GPUs**, including inference-optimised accelerators and emerging thermodynamic co-processors, rather than anchoring procurement exclusively to the current Nvidia GPU ecosystem.

4) Remove obstacles to exploiting datasets particular to sovereign AI applications and challenges, and conduct analysis to determine the benefits and costs of a **data-trading approach** in the UK context.

5) Build a **secure sandbox for developing and testing next-generation AI capabilities**. This should include **shared reference environments and simulators**

[118] Author interview with government participant, 11 February 2026.

[119] Tracy Qu, "How to buy and sell data? Shanghai starts new exchange for trading massive amounts of data like commodities," *South China Morning Post*, 27 November 2021, https://archive.ph/mZZTf; Ann Cao, "Shanghai expands scope of virtual asset trading to become a 'data industry innovation highland' worth US$69 billion by 2025," *South China Morning Post*, 16 August 2023, https://archive.ph/PG6xa.

for high-value government and CNI use cases, so that agentic and autonomous systems can be tested in realistic but controlled settings before deployment.

6) Support **shared access to testbeds and platforms for emerging hardware pathways** (including neuromorphic systems and thermodynamic computing) that allow UK researchers to **evaluate operational relevance, tooling, and integration challenges** before committing to larger-scale investment.

4.3.4 Supporting adoption at scale

As discussed in Section 4.1, harnessing the benefits of AI not only relies on being the first to develop state-of-the-art models, but also on the ability to adopt AI successfully at scale. There must be a stronger focus on downstream adoption, so that frontier and other systems translate into real gains in high-value sectors,[120] such as those identified in the UK's Modern Industrial Strategy.[121] This requires research and policy to identify sector-specific R&D opportunities. Stronger testing, assurance, validation and verification for systems deployed in operational settings where model performance is intended to convert into economic and strategic value is also critical. Survey data found strong support for assurance, governance and safety frameworks that generalise beyond LLMs.

Providing *already world-leading* industries with the tools to benefit from AI is integral to securing further breakthroughs and cementing leverage.[122] In the UK, achieving this requires closer cross-sector partnerships to create shared problem books that connect pioneering research fields to public- and private-sector use cases and data libraries.

The UK Government can also lead by example, becoming an intelligent customer for AI. Alongside making it easier to understand and prioritise government problems, there is a broader need for the Government to be a more consistent customer for UK startups. There should be an initiative to reduce the friction in government adoption via fast-tracked procurement processes, and a 'Challenge List for Government Services' linked to seed funding. These could emulate the National Cyber Security Centre's problem books,[123] and link to the R&D Missions Accelerator Programme.[124]

[120] Author interview with academic participant, 5 February 2026.
[121] HM Government, *Sector Plans* (Department for Business and Trade: September 2025), https://www.gov.uk/government/publications/industrial-strategy-sector-plans/sector-plans.
[122] Author interview with government participant, 4 February 2026.
[123] National Cyber Security Centre, "The NCSC research problem book," last modified 10 July 2024, https://www.ncsc.gov.uk/collection/problem-book.
[124] UK Research and Innovation, "R&D Missions Accelerator Programme," https://www.ukri.org/what-we-do/browse-our-areas-of-investment-and-support/rd-missions-accelerator-programme.

This report proposes the following recommendations for supporting adoption at scale:

1) Support the National Physical Laboratory's new Centre for AI Measurement to evolve into a **stronger government-facing function for testing, monitoring and safe integration** of AI systems in high-priority deployment contexts.

2) Encourage the development and deployment of **specialised lean models for government and other high-priority domains**, where fine-tuned small models can match frontier performance at a fraction of the cost, latency, and data exposure risk.

3) Create a government AI adoption pipeline built around a **Challenge List, shared problem books, curated datasets and test environments, and seed funding for high-value use cases**. Reduce friction through faster procurement routes and lightweight pilot funding for departments that can demonstrate a clear operational use case and assurance plan.

About the Authors

Ardi Janjeva is a Senior Research Associate at the Centre for Emerging Technology and Security (CETaS). His research interests are divided into three main areas: AI innovation and disruption; intelligence tradecraft and investigatory powers; and emerging technology, political economy and strategy. He has worked closely with national and international partners across government, academia, civil society and the private sector on these topics, producing research that has been cited in academic journals and mainstream media outlets such as the Financial Times and the BBC.

Sylvester Kaczmarek is a specialist in advanced AI and cybersecurity for safety-critical cyber-physical systems. He architects secure-by-design, interpretable hybrid AI/ML systems for autonomous space missions and other mission-critical systems, bridging centralised command with distributed edge intelligence. Supported by the UK Space Agency, NCSC, and US Space Force, his work has earned an ESA Award of Merit. He provides strategic expertise to the UK Parliament on AI governance and autonomy, while pioneering foundational frameworks to formally verify non-deterministic and emergent AI behaviours. His efforts set new standards for resilient, trustworthy, off-world robotic intelligence, ensuring the safety and stability of complex cislunar operations.

Angus Shennan is a former Visiting Research Fellow at the Centre for Emerging Technology and Security (CETaS). His research centres on new areas of AI development, foresight mechanisms, and geopolitical implications of technology advancement. He has a particular interest in social-epistemic impacts of AI. He has over 2 years' experience in senior analytical roles in government, primarily focused on interdisciplinary techniques to present evidenced-based analysis to non-expert senior audiences. Before working in central government, he worked in financial technology and for the NHS. Angus holds a BSc in Physics from Durham University and a MSc in Philosophy of Science from the London School of Economics and Political Science.

Dr Carolyn Ashurst is Head of Research at the Centre for Emerging Technology and Security (CETaS), where she oversees projects covering a range of technology and security topics. The Centre's research aims to enable the effective and responsible use of emerging technology within national security, understand global threats and opportunities, and shape policy and governance. Her previous roles include Turing Research Fellow in Safe and Ethical AI at the Alan Turing Institute, Senior Research Scholar at the University of Oxford, and various technical roles related to machine learning and digital systems within government and finance. She holds a PhD in maths from the University of Bath.

Centre for Emerging Technology and Security

RESEARCH REPORT