Advanced Research + Invention Agency ARIA

Mathematics and modelling are the keys we need to safely unlock transformative AI

v1.0

David "davidad" Dalrymple, Programme Director

CONTEXT

This document describes an early opportunity space from which we believe one or more funding programmes can emerge. We've sketched out some of our early thinking to spark your interest, and invite you to imagine relevant potential programmes with us, or suggest new directions.

In tandem, our first emerging programme hypothesis related to this opportunity space has now been published. Read the thesis and register for updates about programme funding opportunities **here**.

An ARIA opportunity space should be

+ important if true (i.e. could lead to a significant new capability for society),
+ under-explored relative to its potential impact, and
+ ripe for new talent, perspectives, or resources to change what's possible.

We don't yet have known technical solutions to ensure that powerful AI systems interact as intended with real-world systems and populations. A combination of scientific world-models and mathematical proofs may be the answer to ensuring AI provides transformational benefit without harm.

BELIEFS

The core beliefs that underpin/bound this area of opportunity.

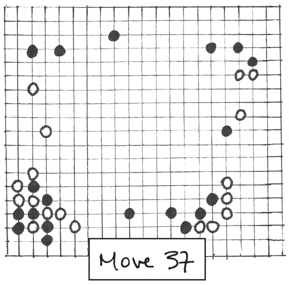
1. Future AI systems will be powerful enough to transformatively enhance or threaten human civilisation at a global scale ⟶ **we need as-yet-unproven technologies to certify that cyber-physical AI systems will deliver intended benefits while avoiding harms**.
2. Given the potential of AI systems to anticipate and exploit world-states beyond human experience or comprehension, traditional methods of empirical testing will be insufficiently reliable for certification ⟶ **mathematical proof offers a critical but underexplored foundation for robust verification of AI**.
3. It will eventually be possible to build mathematically robust, human-auditable models that comprehensively capture the physical phenomena and social affordances that underpin human flourishing ⟶ **we should begin developing such world models today to advance transformative AI and provide a basis for provable safety**.

OBSERVATIONS

Some signposts as to why we see this area as important, underserved, and ripe.

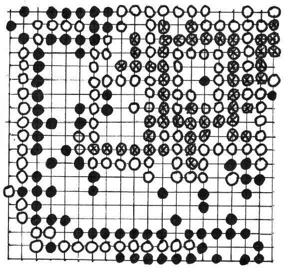
AI holds the potential to dramatically improve physical health, economic well-being, and human empowerment, on a scale exceeding the industrial revolution—if deployed wisely [1].

Leading AI researchers and CEOs have all acknowledged the serious risk that AI systems may cause human extinction, and that "currently, we don't have a solution for steering or controlling a potentially superintelligent AI and preventing it from going rogue" [11, 12, 13, 14].



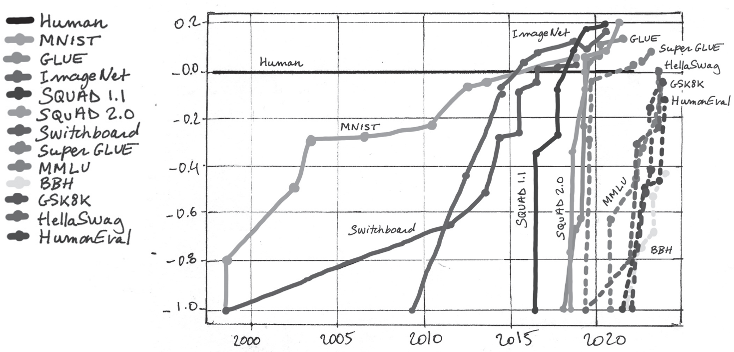
AI systems can exploit states of play beyond human experience or comprehension.

and yet...



Even "strongly superhuman" Go AIs have surprising failure modes, illustrating the limits of benchmarking [15].

Major AI benchmarks and their progress [16]

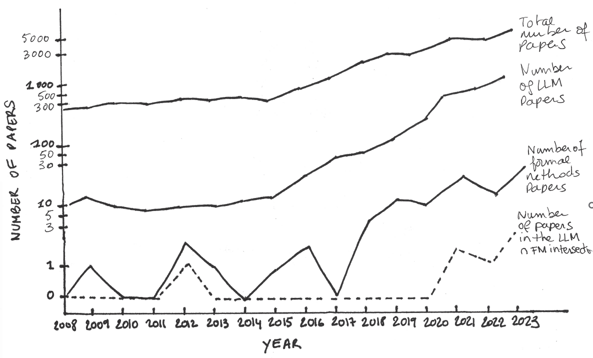


Of papers at top AI conferences, <0.4% mention keywords related to mathematical proof or similar formal methods. Instead, the dominant assessment paradigm by far is benchmarks—which fundamentally rely on statistical assumptions that are only sound in the hypothetical limit of infinite-size test sets.

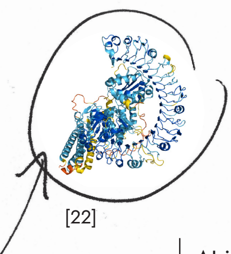
Major frontier AI labs are focused on approaches to alignment and safety [11, 17, 18] which do not target any proof-like guarantees [2, 5, 7-9].

Despite the relative lack of attention, the recent work at this intersection is exciting, much of it becoming feasible only this year with the latest generation of LLMs [3, 4, 19, 20].

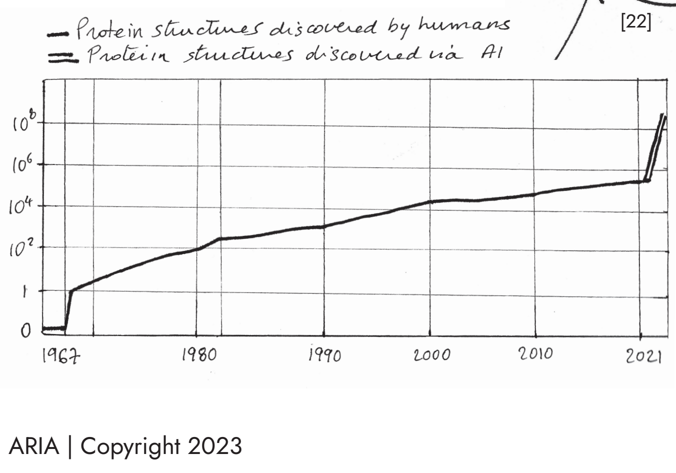
Comparison of (ICML, ICLR, NeurIPS) papers on LLMs vs Formal Methods



Formal methods are increasingly applicable to neural networks, including AI systems larger than 10^8 parameters [6, 21].



[22]



AI is already beginning to enhance the development of scientific world-models relevant to civilisation-scale problems, such as cancer [22] and fusion energy [23].

While formally verifying fully general AI may be impossible, we can likely use it to develop problem specifications, and then certifiable solutions, to ambitious tasks.

ENGAGE

We invite you to shape our efforts by providing feedback and surfacing breakthrough ideas related to this opportunity space. Our next step will be to formulate a programme that directs funding across research disciplines or institutions toward a focused objective. We also plan to open up seed funding for researchers whose bold aspirations are unlikely to be funded elsewhere

Sign up **here** for updates, or to inform the programme thesis. You can upload a short 2 page PDF — we will read anything you send.

If you require an accessible version of this document and/or form, please contact us at info@aria.org.uk.

SOURCES

A compiled, but not exhaustive list of works helping to shape our view and frame the opportunity space (for those who want to dig deeper).

1. The transformative potential of artificial intelligence
2. Provably safe systems: The only path to controllable AGI
3. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics
4. Llemma: An open language model for mathematics
5. Toward verified artificial intelligence
6. Formal verification for neural networks via branch-and-bound
7. COOL-MC: A comprehensive tool for reinforcement learning and model checking
8. Probabilistic model checking and autonomy
9. Automated verification and synthesis of stochastic hybrid systems: A survey
10. Probabilities are not enough: Formal controller synthesis for stochastic dynamical models with epistemic uncertainty
11. Introducing superalignment
12. Statement on AI risk
13. CEO of AI company warns his tech has a large chance of ending the world
14. The CEO of the company behind AI chatbot ChatGPT says worst-case scenario for AI is 'lights out for all of us'
15. Adversarial strategies beat superhuman go AIs
16. Plotting progress in AI (Figure 1)
17. Some high-level thoughts on the DeepMind alignment team's strategy
18. Anthropic's "core views on AI safety"
19. SatLM: Satisfiability-aided language models using declarative prompting
20. From word models to world models
21. VNN-COMP (Verification of Neural Networks COMPetition)
22. Evaluation of AlphaFold on stability of missense variations in cancer (Protein structure)
23. Magnetic control of tokamak plasmas through deep RL

EXTENDED BIBLIOGRAPHY

For an even deeper dive...

24. Robust control for dynamical systems with non-Gaussian noise via formal abstractions
25. AI scientists: Safe and useful AI?
26. Towards autoformalization of mathematics and code correctness: Experiments with elementary proofs
27. A list of core AI safety problems & how I hope to solve them
28. Towards a research program on compositional world-modeling
29. Collective constitutional AI: Aligning a language model with public input
30. xVal: A continuous number encoding for LLMs
31. An overview of catastrophic AI risks
32. GFlowNets for AI-driven scientific discovery
33. When to trust AI: Advances and challenges for certification of neural networks
34. Eureka: Human-level reward design via coding large language models
35. Faster sorting algorithms discovered using deep reinforcement learning
36. Fairness, accountability, transparency, and ethics (FATE)
37. davidad's bold plan for alignment
38. Sam Altman, the man behind ChatGPT, is increasingly alarmed about what he unleashed
39. Trustworthy autonomous system development
40. Misspecification in inverse reinforcement learning
41. Experimental results from applying GPT-4 to an unpublished formal language
42. Fundamental limitations of alignment in LLMs
43. LeanDojo: Theorem proving with retrieval-augmented LLMs
44. Language to rewards for robotic skill synthesis
45. Democratic inputs to AI
46. Neural abstractions
47. Individual fairness guarantees for neural networks
48. Discovering faster matrix multiplication with RL
49. LCRL: Certified policy synthesis via logically-constrained reinforcement learning
50. Provably beneficial artificial intelligence
51. Goal misgeneralization
52. Autoformalization with large language models
53. Learning control policies for stochastic systems with reach-avoid guarantees
54. Advancing mathematics by guiding human intuition with AI
55. The seL4 microkernel: An introduction
56. Safety verification of deep neural networks
57. The basic AI drives