Mathematics for Safe AI

Opportunity space

v1.0

David "davidad" Dalrymple, Programme Director

CONTEXT

This document describes an opportunity space - an area that we believe is likely to yield breakthroughs, from which one or more funding programmes will emerge.

This opportunity space is not currently soliciting feedback — you can stay up to date with this opportunity space, plus others across ARIA, **here**. (www.aria.org.uk/opportunity-space-updates).

In tandem, our programme hypothesis related to this opportunity space has now been published. You can read this document **here**. [PDF] (https://www.aria.org.uk/wp-content/uploads/2024/01/ARIA-Safeguarded-AI-Programme-Thesis-V1.pdf)

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.

SUMMARY

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

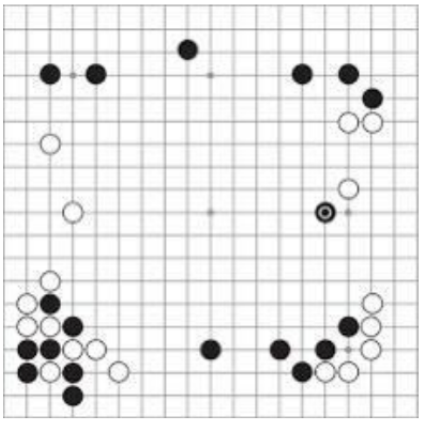
1. Future AI systems will be powerful enough to transformatively enhance or threaten human civilisation at a global scale → we need as-yet-unproven technologies to certify that cyber-physical AI systems will deliver intended benefits while avoiding harms.
2. Given the potential of AI systems to anticipate 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

Sign posts as to why we see this area as important, under-explored, and ripe.

1. 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].

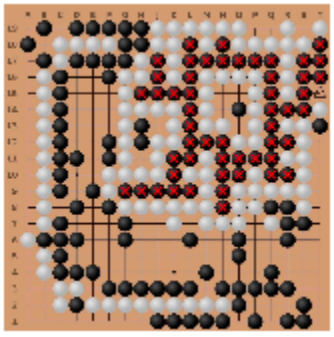
2. Leading AI researchers and CEOs have all acknowledged the serious risk that AI systems may cause human extinction, and that "currently, we don't have a solution for steering or controlling a potentially superintelligent AI and preventing it from going rogue" [11, 12, 13, 14].



3. Figure 1: This image, ‘Move 37’ shows a state of play in the famous game of Go between former human world champion Lee Se-dol and the DeepMind AI system AlphaGo. The 37th stone placed on the board during the game was played by AlphaGo and shocked the professional Go-playing community, as it displayed unprecedented inventiveness.

    Note from davidad — and yet…

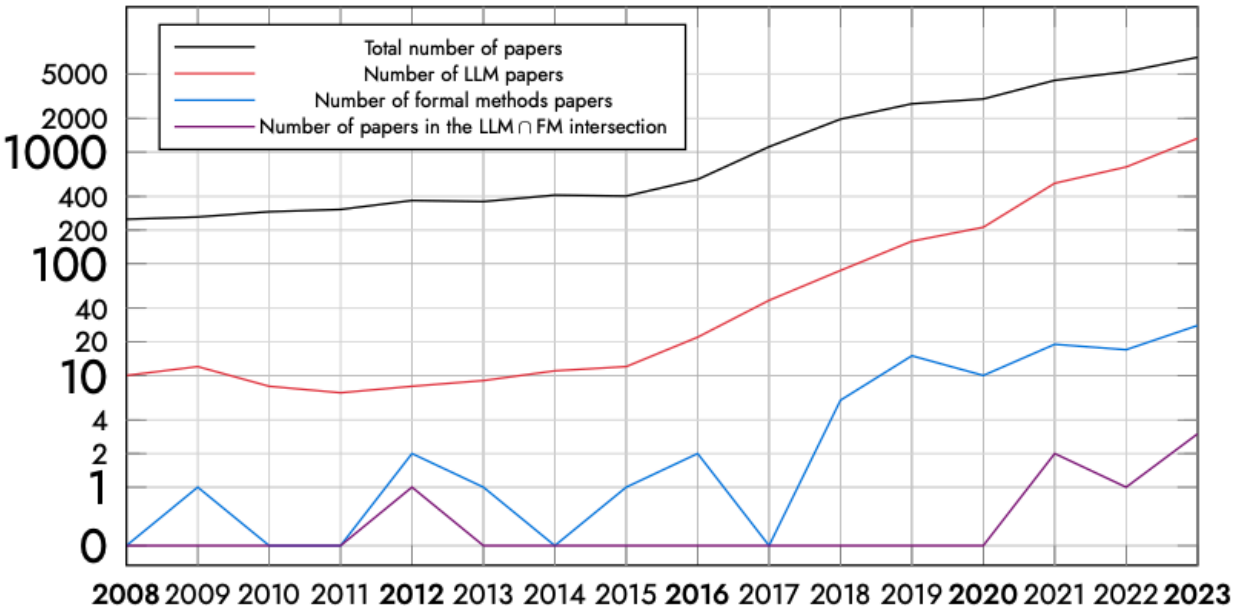
4. AI systems can exploit states of play beyond human experience or comprehension.



5. Figure 2: This image, ‘Cyclic Attack’, shows another Go board, which is much closer to an endgame. This game was played between an AI system as Black and a strong amateur, Kellin Pelrine, as White. Kellin learned an “adversarial attack” in which the AI system fails to recognize when a “cyclic” or doughnut-shaped pattern of its stones is about to be captured, and uses this trick to defeat AI systems considered even stronger than AlphaGo

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

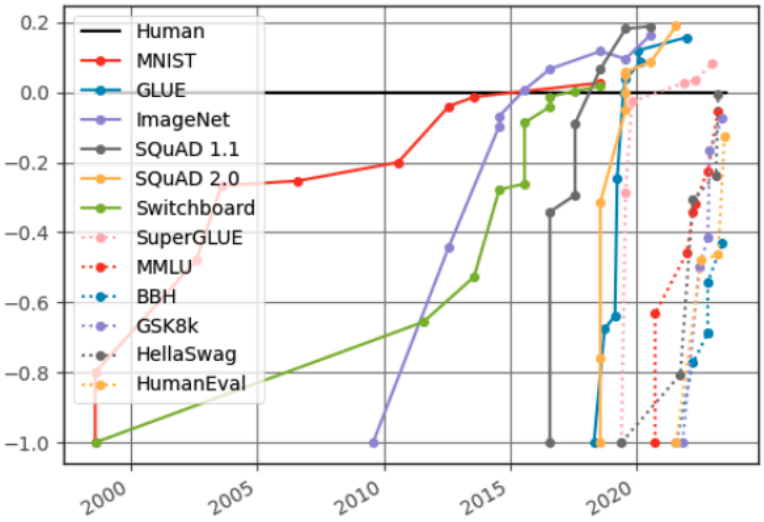
7. Of papers at top AI conferences, <0.4% mention keywords related to mathematical proof or similar formal methods. Instead, the dominant assessment paradigm by far is benchmarks—which fundamentally rely on statistical assumptions that are only sound in the hypothetical limit of infinite-size test sets.



8. Figure 3: This is a graph of the number of papers in various subsets of AI and machine learning research published per year, from 2008 to 2023. The field as a whole has grown exponentially from less than 300 to over 5000 papers per year. Since 2015, there has also been exponential growth in papers regarding Large Language Models in particular, from 10 to over 1000 papers per year. Meanwhile, the trend for formal methods in AI is just getting started, from about 3 papers per year in 2018 to 30 papers per year in 2023.

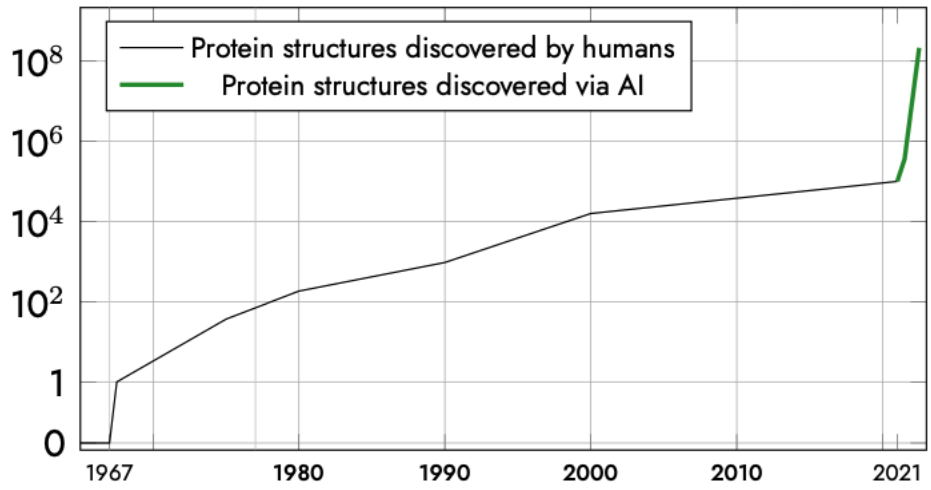
1. Despite the relative lack of attention, the recent work at this intersection is exciting, much of it becoming feasible only this year with the latest generation of LLMs [3, 4, 19, 20].

Note from davidad — 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, 8, 9].

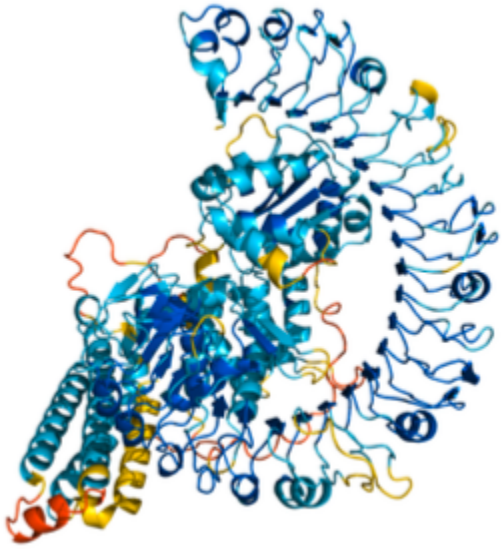


2. Figure 4: The graph shows over a dozen benchmarks of AI performance on tasks such as question-answering and coding relative to typical human performance. Not only have AI systems surpassed human performance on most benchmarks, but the rate at which this happens for new benchmarks appears to be accelerating since 2019.

3. Formal methods are increasingly applicable to neural networks, including AI systems larger than 100 million parameters [6, 21].



4. Figure 5: This graph shows the number of biological protein structures that have been solved over time, from 1967 (when the first protein structure was discovered) to 2023. The AlphaFold project marks a dramatic inflection point in which AI systems have greatly increased humanity's overall pace of discovering new protein structures



5. Figure 6: This image shows an illustrative example of a three-dimensional protein structure.

6. 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].

7. 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

This opportunity space is not currently soliciting feedback – you can stay up to date with this opportunity space, plus others across ARIA, **here**. (www.aria.org.uk/opportunity-space-updates).

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. Provable 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 super alignment
12. Statement on AI risk
13. CEO of AI company warns his tech has a large change 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 (Fig 4)
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 of cancer (Fig 6)
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. Toward 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 programme 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 and 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 skills 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 for deep neural networks
57. The basic AI drives