Safeguarded AI - Atlas Computing's Unaffiliated Summary

"A combination of scientific world-models and mathematical proofs may [ensure] AI provides transformational benefit without harm." - Safeguarded AI programme thesis

BLUF: AI systems are acting with limited safety assurances in increasingly high-risk environments despite being less understood than conventional engineered systems. Safeguarded AI is a £59M ($74M USD) research agenda within UK's ARIA to achieve safety assurance for AI at the same level we expect in conventional safety-critical engineering.

AI experts widely agree: AI will dramatically grow risks to society within years, not decades, and we lack the technology to keep AI safe. Catastrophic risks include terrorist-engineered pandemics, ever-worsening cyberattacks, disruption to power and water supplies, autonomous lethal weapons, extreme power concentration in autocracies or individual corporations, and loss of control over AI systems.[1]

We need quantitative guarantees of critical safety criteria, but current approaches can't provide this. Quantitative guarantees put numbers on risks and identify modeling assumptions, like in the claim "there's a 1-in-a-billion chance this bridge collapses within 10 years if its load stays below 3,000 tons." Governments, businesses, and engineers demand quantitative guarantees on critical safety properties for systems from airplanes to smart phones to medicine, and we need the same norm for AI. Unfortunately, mainstream approaches to measuring or providing safety in modern AI systems are limited to example-based testing (e.g. red-teaming, evals) and training (e.g. RLHF), which by their nature do not generalize to unexpected situations. Multiple factors make this even worse: machine-learning systems are grown, not engineered, raising the prevalence of unanticipated behavior, and research has already demonstrated that vulnerabilities can be hidden during a testing period and only activated at a future time.[2]

A Guaranteed-Safe AI approach is designed to deliver these quantitative safety guarantees. Guaranteed-safe AI approaches make models of what matters, then prove that all safety requirements are satisfied given the modeling assumptions. Since these "world models" are human-auditable, similar to today's simulation software, human domain experts can review and correct world models in detail, and open-sourced world models can achieve a high degree of robustness and community trust. Likewise, safety requirements can be published and reused. Requirements by different stakeholders are easily composed and could be enforced by regulations.

Formal verification brings confidence to verifying safety requirements at scale. Formal verification makes mathematically-sound proofs that can be automatically checked. It is used in industry and military to provide the highest levels of safety assurance in aviation[3], computer chips[4], encryption[5], and more[6]. Importantly, formally verifying AI behavior against trusted world models makes it impossible for an AI to violate the modeled requirements, even if the AI is superintelligent or adversarial.

Guaranteed-safe AI approaches combine AI with formal verification through four steps:

1. Model: Domain experts develop a world model of the domain of concern, e.g. power grids
2. Specify: Users specify goals in terms of the world model, including safety requirements
3. Generate: AI proposes a solution that optimizes the goals and a proof that it satisfies the safety requirements
4. Review: Proof checkers verify that the proposed solution meets the requirements

---

[1] Hendrycks et al. - An Overview of Catastrophic AI Risks

[2] Hubinger et al. - Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training

[3] Fisher et al. - Sparking the New Age of Formal Verification at DARPA

[4] Seligman et al. - Formal verification: an essential toolkit for modern VLSI design

[5] Erbsen et al. - Simple High-Level Code for Cryptographic Arithmetic - With Proofs, Without Compromises

[6] E.g. microkernels (seL4), compilers (CompCert, VST), and transport libraries (WireGuard, Project Everest)

Safety advances will initially come in isolated domains (e.g. cyber, cyber-physical, bio), but these domains can be composed to achieve integrated safety. Computational models, requirements-setting, and correctness arguments are already a standard part of every major discipline in science and engineering and an essential part of modern safety. While a guaranteed-safe AI approach is challenging to achieve, it builds on the existing foundation of organized knowledge.

Advances in AI-assisted formal verification point towards practical applications of provable properties for AI safety within years, not decades. Modern AI systems are great at knowledge-informed search but can't tell if answers are right. This perfectly complements formal methods, which struggle with search but know when answers are right. Recent research advances in proof automation[7], code synthesis[8], and informal-to-formal specification translation ("autoformalization")[9] pave the way to big gains.

UK ARIA's Safeguarded AI program aims to demonstrate a practical workflow for deploying AI agents with guaranteed safety properties to cyber-physical domains. In a successful workflow, trained humans can efficiently construct trusted specifications (i.e. world models and requirements) and manage incremental specification changes. Once confidence is established in a specification, an AI model would be safeguarded to provably obey the specification. Naturally, humans might need to adapt the specification in order for AI-safeguarding to succeed. Upon success, the safeguarded AI model would be deployed to its target domain and would need to demonstrate competitive economic value. While ambitious, achieving these goals is possible and would remove the key technical risks to practical, provable-properties-safe AI.

The Safeguarded AI program will invest £59M into research in many of the biggest hurdles to guaranteed-safe AI. The program is organized into three technical areas (TAs):

1. TA1 tackles world-modeling challenges, including efficient domain-specific modeling, cross-domain interoperability, and model version control. Work on each of these topics is specialized into theory, backend implementation, and HCI contributions.
2. TA2 tackles AI-assisted verification and the production and approval of verified AI. Work focuses on AI aids for world modeling and proof search, techniques for training neural models to verifiably satisfy a specification, and human processes for spec-based oversight and approval.
3. TA3 tackles applications, demonstrating that a safeguarded AI approach can deliver economically-competitive cyber-physical applications.

Incremental successes in guaranteed-safety properties will provide huge windfalls to cybersecurity and other critical infrastructure safety. These windfalls provide fuel for a virtuous cycle of technology improvement and profit. Even without verified neural models, verifying conventional cyber and cyber-physical infrastructure mitigates catastrophic risks from accidents and adversarial attacks. Cybercrime costs are in the high trillions of dollars. Reducing the costs of world-modeling and proof search can unlock tremendous economic and security value.

To go deeper, check out the ARIA Safeguarded AI webpage, including the TA1.1 solicitation presentation and call for proposals. Or, read advocacy for similar approaches from other world leaders in AI safety:

- Yoshua Bengio, 2024 - Towards a Cautious Scientist AI with Convergent Safety Bounds
- Max Tegmark, Steve Omohundro, 2023 - Provably safe systems: the only path to controllable AGI
- Stuart Russell, 2022 - Provably Beneficial AI

---

[7] First et al. - NeurIPS Tutorial on Machine Learning for Theorem Proving

[8] Code Generation on HumanEval

[9] Wong et al. - From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought