Safeguarded AI:

Programme Thesis

v2.0

Nora Ammann, Programme Director

CONTEXT

This document presents the core thesis underpinning a programme that is now launched.

Sign up here to receive all updates about this opportunity space and see the programme here.

An ARIA programme seeks to unlock a scientific or technical capability that

+ changes the perception of what's possible or valuable
+ has the potential to catalyse massive social and economic returns
+ is unlikely to be achieved without ARIA's intervention.

UPDATE: OUR THINKING, EVOLVED

The original Safeguarded AI thesis, published in early 2024, set out to demonstrate that frontier AI can deliver real-world economic impact in business- and safety-critical domains under mathematical safety guarantees — and, by doing so, build consensus that powerful AI should only deploy through verifiable 'gatekeeper' frameworks, reducing the risks posed by advanced AI.

Two things have shifted since. First, the pace of frontier AI progress has moved faster than expected at the time of writing the original thesis (see here and here), which made us decide to relocate the £18m originally earmarked for ML fine-tuning towards an increased commitment into Technical Area 1's tooling efforts, increasing both scope of the ambition

and speed of delivery. Second, timely international consensus on a gatekeeper workflow has become less likely, leading us to focus on accelerating improvements to societal resilience instead. We have discussed these changes here.

This document is a fresh, full articulation of the new programme vision. The original programme thesis is viewable here.

PROGRAMME THESIS, SIMPLY STATED

An overview of the programme thesis, accessible & simply stated

AI agents can now write more software in an hour than human teams used to write in months.[1] But our ability to verify whether that software is correct has not kept pace. This gap has become evident in software engineering; it will soon become equally pressing in a growing number of domains of science, engineering and decision making more broadly.

The programme's ambition is to close this gap by radically transforming what is possible through AI-enabled mathematical modelling and formal verification — expanding its speed, scale, and scope far beyond what is possible today.

If we succeed, societal resilience takes a step-change forward. We can recruit AI to build high-assurance systems in the domains that matter most — secure software and hardware, and critical cyber-physical infrastructure. Organisations can responsibly adopt AI-enabled decision-support and autonomous control in business- and safety-critical domains that demand quantitative assurances of safety, and where the socioeconomic benefits run to hundreds of millions of pounds annually (e.g. energy grid balancing, supply chain management, clinical trial design). And we can leverage AI in scientific modelling and decision support without ceding control to unexplainable 'black boxes'.

[1] See, for example, MirrorCode, where Claude Opus 4.6 autonomously reimplemented a 16,000-line bioinformatics toolkit estimated at 2–17 weeks of human engineering effort.

To carry this from vision into reality, we invest in the theoretical foundations, software infrastructure, human-AI collaboration tooling, and real-world applications.

- **Technical Area 1 — Tooling** builds an open-source mathematical assurance toolkit that lets fleets of AI agents produce formally verified artifacts at unprecedented speed and scale. This work breaks down into four guiding questions:
  - **TA1.1 — Theory**: How can we expand the types of systems we can formally model, express safety-relevant specifications about and synthesise proof certificates for?
  - **TA1.2 — Infrastructure**: How can we model, store, query and analyse formal artifacts efficiently?
  - **TA1.3 — Interaction Paradigms**: How can humans and AI collaborate effectively on authoring and reviewing formal artifacts?
  - **TA1.4 — Socio-Technical Integration** How can we ensure the resulting capabilities are beneficial, trustworthy and governable?
- **Technical Area 2 — Applications** stress-tests and demonstrates these capabilities in high-value, societally critical application domains, covering cyber- and cyber-physical systems.

If we succeed, the most consequential applications won't be the ones we demonstrate within the programme's lifetime; they will be the ones a growing ecosystem builds after the programme ends.

In a nutshell...

TA1 - Tooling	
TA1.1 - Theory What systems and behaviours are we able to model formally?	**TA1.2 - Infrastructure** How can we model, store, query and analyse formal artifacts efficiently?
TA1.3 - Interaction Paradigms How can humans and AI collaborate effectively on authoring and reviewing formal artifacts?	**TA1.4 - Socio-Technical Integration** How can we ensure the resulting capabilities are beneficial, trustworthy and governable?

TA2 - Applications
How can we demonstrate and apply these capabilities in societally critical domains? *E.g. software verification, cyber-physical control, hardware security, scientific modelling & decision support, etc.*

This programme thesis lives in the ARIA opportunity space: Mathematics for Safe AI.

PROGRAMME THESIS, EXPLAINED

A detailed description of the programme thesis, presented for constructive feedback

Why this programme

The world is speeding up. AI itself is accelerating, and it accelerates and transforms most other domains of science, technology, and our social, economic, and political lives in turn. The bottleneck, increasingly, is not just raw AI capabilities, but whether we can direct these capabilities toward outcomes we can trust.

AI systems can already code at superhuman levels in narrow settings, and that capability is extending into the physical world, our scientific practices, and increasingly consequential areas of decision-making. As the volume and complexity of AI-generated artefacts outpace the engineering and auditing workflows we built around human authors, we stand to pay the cost twice: in security incidents we fail to prevent, and in beneficial AI applications we fail to adopt because the trust is not there.

By radically upgrading our capabilities in formal modelling and verification, we can unlock a new space of possibilities of what we can do with advanced AI. We envision a world where AI-produced artifacts — including code, microarchitectures, cryptographic primitives, scientific models, decision support and cyber physical control systems — carry quantitative guarantees of correctness and fail-safety. In this world, AI can be trusted to operate in domains that today require human-in-the-loop oversight at every step, and where it can operate faster, over larger systems, with stronger guarantees than a human team alone could ever achieve.

What we are trying to build

The programme is structured around two main technical areas. TA1 makes long-horizon technical bets on the theoretical and infrastructural foundations that, if successful, would transform the reach of AI-enabled formal methods. In TA2, we fund teams to apply current capability to consequential real-world problems, pushing the frontier of what is practically possible.

The programme has been underway since mid-2024 and is due to conclude at the end of 2027. At the time of writing (May 2026), Technical Area 1 (Tooling) is fully funded, with additional funds yet to be allocated in Technical Area 2 (Applications).

TA1.1 — Theory.

To reason about, analyse, or verify a system, we first need to model it.

The systems this programme cares about cut across modelling traditions. Real cyber and cyber-physical systems mix discrete control with continuous dynamics, deterministic computation with stochastic environments, classical components with neural ones. A wide range of domain-specific modelling languages already exists (e.g. differential equations, Markov processes, Petri nets, probabilistic graphical models, hybrid automata). But hard-coding any single one would prematurely cap the scope of potential applications.

Our aim in TA1.1 is therefore to determine a **meta-ontology** that acts as a common substrate for hybrid, compositional modelling, in which application-specific modelling languages can be expressed, combined, and refined as needed. The substrate must be expressive enough to capture the systems we care about, structured enough that artefacts can be efficiently reasoned about and co-developed by AI agents and human domain experts. Concretely, we want to transcend assumptions that state spaces be finite, discrete, or finite-dimensional (constraining only to σ-compact); and to transcend assumptions about the type of dynamics (deterministic, stochastic, nondeterministic, graphical, temporal); within an epistemic framework that hosts both Bayesian and Knightian uncertainty. Appendix 2 contains an illustrative list of real-world systems that we want to be able to formally model.

Our bet on what makes this achievable is categorical and diagrammatic foundations. Graphs, ASTs, IRs, Petri nets, string diagrams, hybrid automata, and electronic circuits are all diagrams in the categorical sense: collections of objects related under specific rules. That commonality is what lets a single substrate host all of them, and what makes compositional reasoning actually work.

Beyond a model of the system's dynamics, we also need to express and reason over the constraints we want the dynamics to satisfy and the evidence that they do. **Specifications** capture which behaviours are acceptable, as probabilistic temporal-logic claims and counterfactual queries, e.g. in the form of an upper bound on the probability of a harmful outcome relative to a do-nothing counterfactual. **Certificates** are the evidence that a system meets its specs, from classical (barrier functions, abstract interpretation, branch-and-bound, Alethe / LFSC certificates from SMT solvers) to neural (reach-avoid supermartingales, control barrier functions). Finally, **neural systems** themselves must also be expressible because the specifications we want to verify refer to systems that contain neural networks (e.g. a neural controller). This is less a semantic challenge than a practical one: a neural network is a continuous function, which the substrate already accommodates, but we want it to expose tensor structure so storage and inference algorithms can exploit it.

We aim to deliver a reference definition of these languages: formal semantics, composition operators, a serialisable structure that supports version control, along with a library of composability results. In its most ambitious form, we are pursuing something like Codd's 1970 relational model and the SQL ecosystem that grew on top of it: a new foundation for representing uncertain, compositional knowledge, suited to an age in which the heaviest users are AI agents rather than humans.

For further detail, see the (now closed) TA1.1 call for proposals (here) and its accompanying presentation (here).

TA1.2 — Infrastructure.

Theory determines what we can express formally. The backend defines what AI agents can *do* with those formal artifacts, at scale.

Today's proof infrastructure was designed for humans. It is not designed for AIs working in parallel to efficiently reason over formal artifacts and proofs potentially several orders of magnitude larger than what has been done so far. To unlock such scale, we need to redesign the backend around how AI agents work. Our current thinking centres on the following high-level desiderata:

- *Machine-native representation*. Programs, specifications, and proofs are stored as structured semantic objects (AST/IR-like) rather than source files. Decomposition, querying, and manipulation of large world-models becomes tractable in ways text doesn't permit, and LLMs can interface directly with the underlying structure rather than through a flattened textual representation.
- *Versioned, structured collaboration*. Native branching, backtracking, and merging — with merges happening over structured semantic objects, not text. Branches, dead ends, and successful paths are preserved as provenance; merge conflicts identify the violated constraint, not just a textual diff.
- *Proof-aware writes*. Database constraint checking and proof checking become the same operation. Writes can be conditional on accompanying proofs; any theorem stored is guaranteed true (modulo the proof checker and its axioms). Constraint checking is incremental, so adding new facts doesn't force re-checking the entire database.
- *Fast, parallel checking*. The checker sits close to the model, potentially co-located with the LLM driving generation. The architecture is parallel, not gated on a serial kernel bottleneck, so the propose/check loop is tight enough to support agent-paced iteration.
- *Local-first execution*. Each agent reads and writes its own replica; writes replicate asynchronously. Many agents can explore proof paths in parallel without blocking each other or being gated on a network round-trip.

Beyond classical proof checking, the backend needs to support the kinds of proof search that the most consequential systems demand. Cyber-physical systems with neural controllers have state spaces too large and dynamics too entangled for techniques that were designed around discrete, hand-written programs. Finding a closed-form proof that a neural network controlling a power grid keeps it stable, or that a complex concurrent software system preserves its invariants across all possible execution orders, may not be tractable. Neural certificates are a promising alternative approach: a compact mathematical object whose existence implies the property of interest, and whose validity can be checked far more cheaply than it can be found. Lyapunov functions are the canonical example; their modern, learned cousins extend the same idea to systems where writing the certificate by hand is hopeless but training one is tractable. The hard work of search is offloaded to AI, while the checker stays small and trusted. For this to be infrastructure rather than a one-off, the backend has to treat certificates as first-class citizens: storing them alongside the world-models they certify, retrieve counterexamples when they fail, and feed both back into the next round of agentic search.

Our goals are ambitious, but we have decades of progress to build on: distributed databases (local-first replication, asynchronous merging, columnar storage at scale), version control (CRDT-style branching and provenance), proof assistants (small trusted kernels, dependent type theory, the discipline of separating trusted from untrusted code), compiler intermediate representations (structured, machine-readable representations of programs at multiple levels of abstraction), and category theory (compositional, diagrammatic foundations for combining heterogeneous models). Each strand has matured substantially in isolation. The bet of TA1.2 is that integrating them around AI agents as the primary target user is now both possible and timely.

For further detail, see the (now closed) TA1.2 call for proposals (here) and its accompanying presentation (here).

TA1.3 — Human-AI interaction.

Theory and infrastructure let AI agents produce formal artifacts at scale. TA1.3 builds interaction paradigms that let humans steer and audit this work.

As proof search and verification scale up, the trust base concentrates into the specs and surrounding assumptions — exactly the artifacts humans need to understand and audit carefully. TA1.3 explores new interaction paradigms that make this possible at scale. The interaction layer's role is to surface the decisions humans need to make and to keep the rest of the work legible enough to be reviewed thoroughly. AI is also opening up a genuinely new design space for human-computer interaction. Within this design space, and given the specific challenges of AI-enabled R&D, the following are capabilities we're interested in seeing developed:

- **Authoring** — letting human-AI teams interactively co-develop world-models, specifications, and complex R&D plans, moving fluidly between informal intent and formal artifacts, and between textual and diagrammatic representations.
- **Elicitation** — helping stakeholders articulate and reflect on goals, specifications, assumptions, and requirements, including when knowledge is tacit, partial, or held across multiple parties.
- **Auditing** — letting humans build justified confidence in the specifications and assumptions where assurance claims ground out, by surfacing edge cases, counterexamples, hidden assumptions, etc.
- **Scalable oversight** — letting humans maintain a grounded understanding of an R&D process moving orders of magnitude faster than they're used to: directing attention to where it's most needed and making the provenance of specific choices and changes inspectable.
- **Iteration** — designing workflows that let human-AI teams develop, refine, and update plans, models, specifications, and guarantees as they learn more, in ways that match specific contexts of use.

- **Run-time monitoring** — adding defense-in-depth by checking whether incoming observations are consistent with the mathematical model of the environment, and surfacing potentially safety-relevant anomalies as they emerge.

If TA1.3 succeeds, a small human team can reliably steer, audit, and trust the work of a much larger agent fleet. The interaction layer is what carries the weight of legibility and accountability across the gap in throughput between humans and agents.

For further detail, see the (now closed) TA1.3 call for proposals (here) and its accompanying presentation (here).

TA1.4 — Sociotechnical integration.

How do we ensure the resulting capabilities and outputs are trustworthy, governable and deployed in service of humanity at large?

Beyond the technical core, several open challenges are fundamentally socio-technical, requiring interdisciplinary expertise across the social sciences. Some of our leading concerns:

- **Preference elicitation and aggregation.** Specifications reflect choices about, for example, acceptable risk levels or tradeoffs between competing preferences. How do we ensure that these choices are duly informed by and legitimate with respect to the stakeholders they affect? Building on a rich literature in social choice, mechanism design, economics, psychology, pedagogy and beyond, we are interested in processes that let diverse groups of stakeholders deliberate over acceptable risks and safety specifications, and in quantitative bargaining solutions for aggregating such preferences, including in cases (reflective of real-world use cases) where information about such risk attitudes and preferences is partial or partially uncertain.
- **Artifact portability, trust, and privacy.** A proof is only worth as much as the set of parties willing to act on it. For verified artefacts to do real institutional work (e.g. a regulator accepting a vendor's safety case, or one team building on another's verified component), they need to travel across organisational boundaries with their assurance claims intact. A party who didn't produce the verification needs good

reason to trust that the proof actually pertains to the system in question, and not some other system. In some contexts, privacy might be paramount, requiring ways to produce independently checkable proof certificates without exposing sensitive information to third parties. Existing work in fields like zero-knowledge proofs, confidential computing, supply-chain security all bear on these questions, but adapting this work to the scale and shape of the problem at hand requires further work.

- **Stewardship.** The important question is not just whether the technology works, but whether the institutions around it can absorb and govern it without losing the channels through which humans currently keep those institutions broadly aligned with their interests. What institutional mechanisms can sustain accountability, auditability, contestability and ultimately legitimacy in these and related technological capabilities once deployed? How can we ensure the benefits of AI-enabled R&D are broadly shared and the downsides actively mitigated rather than left as externalities?

If TA1.4 succeeds, the specifications our toolkit verifies systems against reflect something defensibly close to what affected publics would endorse on reflection, and the institutions consuming verified artefacts have the legitimacy and accountability structures to act on them.

For further detail, see the (now closed) TA1.4 call for proposals (here).

Technical Area 2 — Applications

The goal is not just to build novel capabilities, but to carry those capabilities into the real world, demonstrate their potential, and catalyse real-world impact.

We target application domains that address problems with the potential for large-scale social or economic impact at the order of hundreds of millions of pounds per year in mature deployment; and insofar as they reveal or push the current frontier of AI-enabled formal methods. We want to understand what kinds of systems can be represented, what scale of artefact can be verified, and how close verified outputs can get to real-world use.

Conceptually, we seek to push that frontier on two axes:

+ **Scope** is what kinds of systems we can formally model and verify: moving from discrete to continuous state spaces; from deterministic to probabilistic to imprecise probabilistic models. from single-process programs to true concurrency/ from pure software to systems whose behaviour depends on microelectronics, physical dynamics, stochastic environments, and learned components. Increasing the scope means expanding the kinds of real-world phenomena that can be brought under precise, checkable formal representation.
+ **Scale** is how close verified artefacts get to deployment: moving from toy examples and isolated lemmas to verified components, subsystems, system-level properties, maintained codebases, and artefacts that survive real-world change, integration, and adversarial pressure.

Our initial focus is cybersecurity. This is an area where we see both urgent need and unusually strong potential to expand what is possible. DARPA's HACMS programme showed this was possible in principle: formally verified components, such as the seL4 microkernel, held up under sustained red-teaming on real systems, including a manned helicopter. But proof effort has historically scaled poorly with system size — and that is exactly what AI-enabled tooling stands to change. AI's growing cyber-offensive capabilities dramatically increase the urgency to help AI-enabled methods mature into a real-world, scalable solution.

As the underlying tooling and capabilities mature, we seek to push into increasingly rich classes of systems: concurrent and distributed software; code verified against hardware and microelectronics models; embedded systems; and cyber-physical systems where verified artefacts support decision-making, automation, or control in domains such as energy, transport, telecommunications, supply chains, industrial processes, R&D planning, and medical devices.

By cyber-physical applications, we mean systems where software, data, and decisions interact with physical or operational processes. In some cases, the output may be certified decision-support, e.g. a planning tool whose recommendations come with checkable guarantees relative to a formal model. In others, it may be a more automated or closed-loop

system consisting of a main controller, runtime monitor, and backup controller whose composite behaviour is certified against safety, robustness, and performance specifications.

Another important application frontier is scientific modelling. Here the goal is not certified autonomy, but to make complex scientific models more formal, compositional, auditable, and useful for scientific progress and high-stakes decision-making, e.g. in climate and weather modelling, epidemiology, ecological and biodiversity modelling, macroeconomic modelling, etc. These domains involve heterogeneous evidence, uncertain causal structure, which may require richer representations of uncertainty and shifting the verification target from correctness to consistency.

AI capabilities are progressing at substantial but uncertain rates, so it remains hard to forecast how far the frontier will expand. We will adopt a dynamic stance, regularly recalibrating what counts as the new ambitious frontier.

A longer list of potential application domains can be found in Appendix 2.

How we think about impact

Programme success, in the narrow sense, means producing credible proof points during the programme's lifetime: TA2 applications that demonstrate AI-enabled formal methods on consequential systems, and a TA1 toolkit that makes those demonstrations possible. The deeper ambition is for those demonstrations, and the tooling behind them, to catalyse a world in which formal modelling and verification become cheaper, more reusable, more trusted, and applicable across a widening set of domains. The most consequential applications, if we succeed, will be the ones a growing ecosystem of researchers, developers, and adopters builds after the programme has concluded.

We aim for three impact pathways:

- **Direct impact.** TA2 produces verified artefacts in the application domains we fund, starting with cybersecurity. These are the most measurable impacts within the programme's lifetime: concrete systems, components, workflows, or assurance cases that real users can evaluate against existing baselines.

- **Ecosystem impact.** TA1's open-source toolkit becomes infrastructure that researchers, startups, labs, and adopters can build on after the programme. If the cost of formal modelling and verification falls, and if the set of systems formal methods can reach expands, the most important applications may be developed by others after the programme has concluded.
- **Discursive impact.** The programme produces evidence — concrete, deployed, quantitatively guaranteed systems — that shifts what the global AI safety, engineering, governance, and standards communities believe is achievable. This kind of impact compounds slowly but durably: it changes the perceived frontier, and therefore changes what others attempt.

To support these impact pathways, work funded under TA1 is open-source or permissively licensed by default, fostering reuse, scrutiny, interoperability, and a community of practice around the toolkit. TA2 applications may have proprietary, sensitive, or commercialisable components depending on the domain, and commercialisation may be effective in driving adoption, attract further investment, and carry verified artefacts into real operational settings.

What we are still trying to figure out

The programme is built on a set of bets about a moving target. The following are open questions we expect to keep revisiting as the technology and the ecosystem evolve.

- Our approach assumes AI agents can substantially uplift formal-methods workflows without needing to be blindly trusted. However, exactly how much leverage we can get from AI still depends on how aligned and reliable these systems are. We therefore continue to monitor how well alignment methods continue to generalise across substantial AI capability advances or paradigm changes, and whether we should assume frontier models remain reasonably steerable through 2030.

- How fast will we be able to expand the scale and scope of AI-enable formal methods, from cybersecurity into other domains of engineering and decision support? Where will the ambitious frontier sit in 2026, 2027, and so on? Will this warrant opening another application-focused funding call in 2027?

- How can we best foster post-programme impact? For TA1, that means working out how to steward an open-source commons: maintenance, documentation, standards, governance, and a community of contributors. For TA2, it means understanding how verified artefacts travel into real operational settings, through commercialisation, domain partnerships and otherwise, without weakening the assurance and trust story that makes them valuable.
- If we are successful, what new challenges and opportunities come into view beyond the programme's immediate scope? How do we ensure this investment does not end as a narrow technical result, but becomes a stepping stone toward the next generation of work enabling resilient, trustworthy, and broadly flourishing futures?

ENGAGE

You can learn about our currently funded projects here. Click here to register your interest, hear about future funding opportunities, or to provide feedback that can help improve this programme thesis.

FURTHER READING

Selected works that motivate, ground, or extend the programme's thinking.

1. Epoch AI, "Epoch Capabilities Index (ECI)," 2025. Available: https://epoch.ai/benchmarks/eci
2. T. Adamczewski, D. Rein, D. Owen, and F. Brand, "MirrorCode: Evidence that AI can already do some weeks-long coding tasks," Epoch AI, Apr. 10, 2026. Available: https://epoch.ai/blog/mirrorcode-preliminary-results
3. T. Hubert et al., "Olympiad-level formal mathematical reasoning with reinforcement learning," Nature, Nov. 2025, doi: 10.1038/s41586-025-09833-y. Available: https://www.nature.com/articles/s41586-025-09833-y
4. Z. Z. Ren *et al.*, "DeepSeek-Prover-V2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition," arXiv:2504.21801, Apr. 2025. Available: https://arxiv.org/abs/2504.21801

5. D. Dalrymple et al., "Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems," arXiv:2405.06624, May 2024. Available: https://arxiv.org/abs/2405.06624

6. B. Fong and D. I. Spivak, *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge, UK: Cambridge University Press, 2019. Available: https://arxiv.org/abs/1803.05316

7. Topos Institute, "CatColab: A collaborative environment for formal, interoperable, conceptual modeling." [Online]. https://catcolab.org/

8. A. Abate, D. Ahmed, M. Giacobbe, and A. Peruffo, "Formal synthesis of Lyapunov neural networks," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 773–778, Jul. 2021, doi: 10.1109/LCSYS.2020.3005328. Available: https://arxiv.org/abs/2003.08910

9. M. Giacobbe, D. Kroening, A. Pal, and M. Tautschnig, "Neural model checking," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Available: https://arxiv.org/abs/2410.23790

10. Đ. Žikelić, M. Lechner, A. Verma, K. Chatterjee, and T. A. Henzinger, "Compositional policy learning in stochastic control systems with formal guarantees," in Advances in Neural Information Processing Systems (NeurIPS), 2023. Available: https://arxiv.org/abs/2312.01456

11. M. Capucci and D. J. Myers, "Compositionality of Lyapunov functions via assume-guarantee reasoning," arXiv:2604.03017, Apr. 2026. Available: https://arxiv.org/abs/2604.03017

12. N. Amit, S. Goldwasser, O. Paradise, and G. Rothblum, "Models that prove their own correctness," in Advances in Neural Information Processing Systems (NeurIPS), 2025. Available: https://arxiv.org/abs/2405.15722

13. "Automerge". 2025. Available: https://automerge.org/

14. S. K. Lahiri, "Intent formalization: A grand challenge for reliable coding in the age of AI agents," arXiv:2603.17150, Mar. 2026. Available: https://arxiv.org/abs/2603.17150

15. K. Fisher, J. Launchbury, and R. Richards, "The HACMS program: using formal methods to eliminate exploitable bugs," Philosophical Transactions of the Royal Society A, vol. 375, no. 2104, Oct. 2017, doi: 10.1098/rsta.2015.0401. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5597724/

16. X. Leroy, "Formal verification of a realistic compiler," Communications of the ACM, vol. 52, no. 7, pp. 107–115, Jul. 2009, doi: 10.1145/1538788.1538814. Available: https://xavierleroy.org/publi/compcert-CACM.pdf

17. G. Klein et al., "seL4: Formal verification of an OS kernel," in Proc. ACM SIGOPS 22nd Symp. Operating Systems Principles (SOSP), Big Sky, MT, USA, Oct. 2009, pp. 207–220, doi: 10.1145/1629575.1629596. Available: https://www.sigops.org/s/conferences/sosp/2009/papers/klein-sosp09.pdf

APPENDIX

1. Context on Naming change

In V1 of the programme thesis, Technical Areas had the following indexing:

TA1: Scaffolding
TA2: ML
TA3: Applications

As a result of the strategic pivot, we cancelled TA2, and refocused the scope of TA3 applications to include cyber-only applications. In V2 of the programme, the indexing has changed to:

TA1: Tooling
TA2: Applications

2. Example application domains

- **Cyber, software, and formal software engineering**
  - end-to-end correctness for cryptographic libraries and protocols
  - proof-checked operating-system kernels and microkernels
  - compiler verification
  - distributed protocol correctness
  - smart-contract and blockchain consensus invariants
  - software supply-chain assurance
- **Concurrency, distributed systems, and parallelism**
  - multicore memory-model conformance
  - lock-free data structures and concurrent algorithms
  - real-time scheduling with worst-case execution-time bounds
  - replicated systems, consensus, sharding, and transaction managers
- **Hardware, microelectronics, and the hardware/software boundary**

    - processor and microarchitecture verification
    - cache coherence and memory models
    - ISA conformance
    - RTL designs verified against architectural specifications
    - firmware verified against hardware or microelectronics models
- **Embedded and cyber-physical control systems**
    - power electronics firmware
    - medical-device controllers
    - automotive ECUs
    - industrial process control
    - robotics in human environments
    - aircraft and spacecraft flight dynamics
    - water and wastewater operations
- **Operational decision-support and optimisation**
    - energy-system dispatch, forecasting, grid modelling, and transmission planning
    - transport routing, signalling, fleet planning, and airspace management
    - telecommunications network optimisation
    - supply-chain and inventory management
    - vaccine and drug supply chains
    - construction project management and scheduling
    - complex healthcare delivery and triage workflows
- **Scientific and public-interest modelling**
    - epidemiology and intervention modelling
    - climate and weather prediction
    - clinical-trial optimisation
    - bioreactor monitoring and tuning
    - R&D planning and roadmapping
    - risk assessment and risk management
- **Data integration and knowledge infrastructure**
    - high-assurance data integration where errors are costly
    - health-record integration
    - scientific or engineering knowledge-base construction
    - model updating from heterogeneous evidence sources