Advanced Research + Invention Agency ARIA

Safeguarded AI: constructing safety by design

Programme thesis

v1.0, David "davidad" Dalrymple, Programme Director

CONTEXT

This document presents the core thesis underpinning a programme that is currently in development at ARIA. We share an early formulation and invite you to provide feedback to help us refine our thinking.

This is not a funding opportunity, but in most cases will lead to one. Sign up here to learn about any funding opportunities derived or adapted from this programme thesis.

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

PROGRAMME THESIS, SIMPLY STATED

This programme thesis is derived from the ARIA Opportunity Space: **Mathematics and modelling are the keys we need to safely unlock transformative AI**.

Imagine a future where advanced AI powers breakthroughs in science, technology, engineering, and medicine, enhancing global prosperity and safeguarding humanity from disasters—but all with rigorous engineering safety measures, like society has come to expect from our critical infrastructure. This programme shall prototype and demonstrate a toolkit for building such safety measures, designed to channel any frontier AI's raw potential not only responsibly, but legibly and verifiably so.

This programme envisions a pathway to leverage frontier AI itself to collaborate with humans to construct a "gatekeeper": a targeted AI whose job is to fully understand the real-world interactions and consequences of an autonomous AI agent, and to ensure the agent only operates within agreed-upon safety guardrails and specifications for a given application. These safeguards would not only reduce the risks of frontier AI and enable its use in safety-critical applications, they would also unlock the upside of frontier AI in *business*-critical applications and commercial activities where reliability is key (unlike entertainment, media, advertising, and sales).

At the end of the programme, we aim to show a compelling proof-of-concept demonstration, in at least one narrow domain, where AI decision-support tools or autonomous control systems can improve on both performance and robustness versus existing operations, in a context where the net present value attainable by full deployment is estimated to be billions of pounds. Some examples of potential such early demonstration areas include: balancing electricity grids, supply chain management, clinical trial optimisation, and 5G beamforming/subchannel allocation for mobile telecommunications networks.

If successful, this would in turn produce a scientific consensus that "AI with quantitative safety guarantees" is a viable R&D pathway that yields key superhuman capabilities for managing cyber-physical systems, **unlocking positive economic rewards**—while *also* building up large-scale civilisational resilience, thereby **reducing risks** from humanity's vulnerability to potential future "rogue AIs"[5,22] to an acceptable level within an acceptable time frame.

PROGRAMME THESIS, EXPLAINED

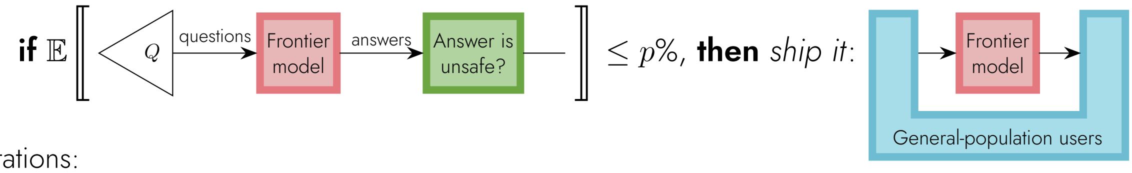
A detailed description of the programme thesis, presented for constructive feedback.

Why this programme

As artificial intelligence becomes exponentially more capable, it has the potential to dramatically improve physical health, economic well-being, and human empowerment, on a scale exceeding the industrial revolution—if deployed wisely[18]. But the current AI development pathway poses severe risks: leading AI researchers and CEOs have all acknowledged that "mitigating the risk of [human] extinction from AI should be a global priority, alongside other societal-scale risks like pandemics and nuclear war"[21].

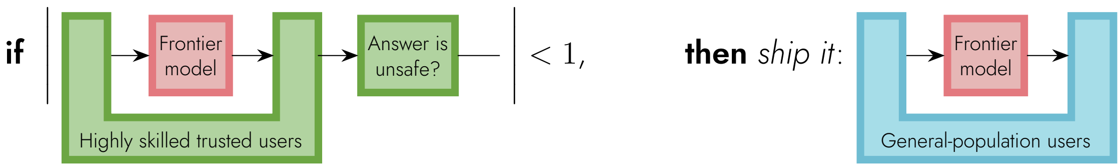
Current work on mitigating this risk is focused primarily on a set of techniques which aim to keep the structure and interface of pretrained frontier AI models intact, while making them safer. These techniques deliver incremental benefits, but have serious limitations, and empirically cannot be relied upon to ensure safety, even in combination[23]. To illustrate, two central examples are:

+ **Evals**, which comprise a finite set Q of "questions" (also known as "prompts" or input strings) on which the evaluator examines a Monte Carlo sample of the frontier AI model's "answers" (also known as output strings) and thereby estimates a propensity of unsafe behaviours, to be quantified before deployment:



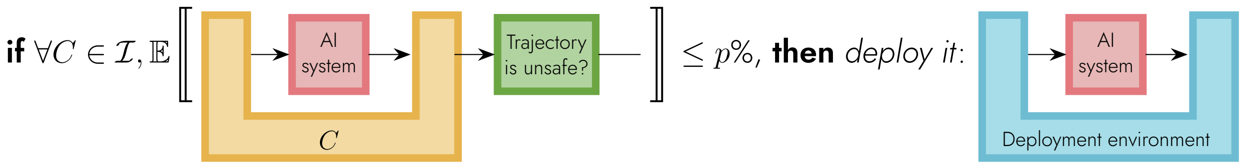
  Limitations:
  - One can obtain confidence that there is a safety problem by uncovering one in an eval, but if no problems are uncovered, this provides very little confidence about whether a safety problem could still emerge if a user employs alternative prompting strategies that are not represented in Q.
  - Users can invoke deployed models in complex recursive ways ("scaffolding"), which greatly expands the space of possible operating conditions that are not checked in an eval. Even if an advanced eval does test *some* scaffolds, the space of potential scaffolds is even more exponential than input strings.

+ **Red-teaming**, in which groups of highly skilled users of AI are tasked with evoking the most unsafe possible outputs from the model, and if they can't find any problems, then the model can be deployed:



  Limitations:
  - Since it involves in-depth interaction with humans, red-teaming is not very scalable.
  - Although humans can exercise some ingenuity in surfacing problems, still, fundamentally, they cannot try everything, and the red-teaming exercise gives very little assurance of what the model might do outside the test coverage.

We would prefer a probabilistic guarantee that universally quantifies over an infinite family $\mathcal{I}$ of plausible initial conditions C of the deployment environment:



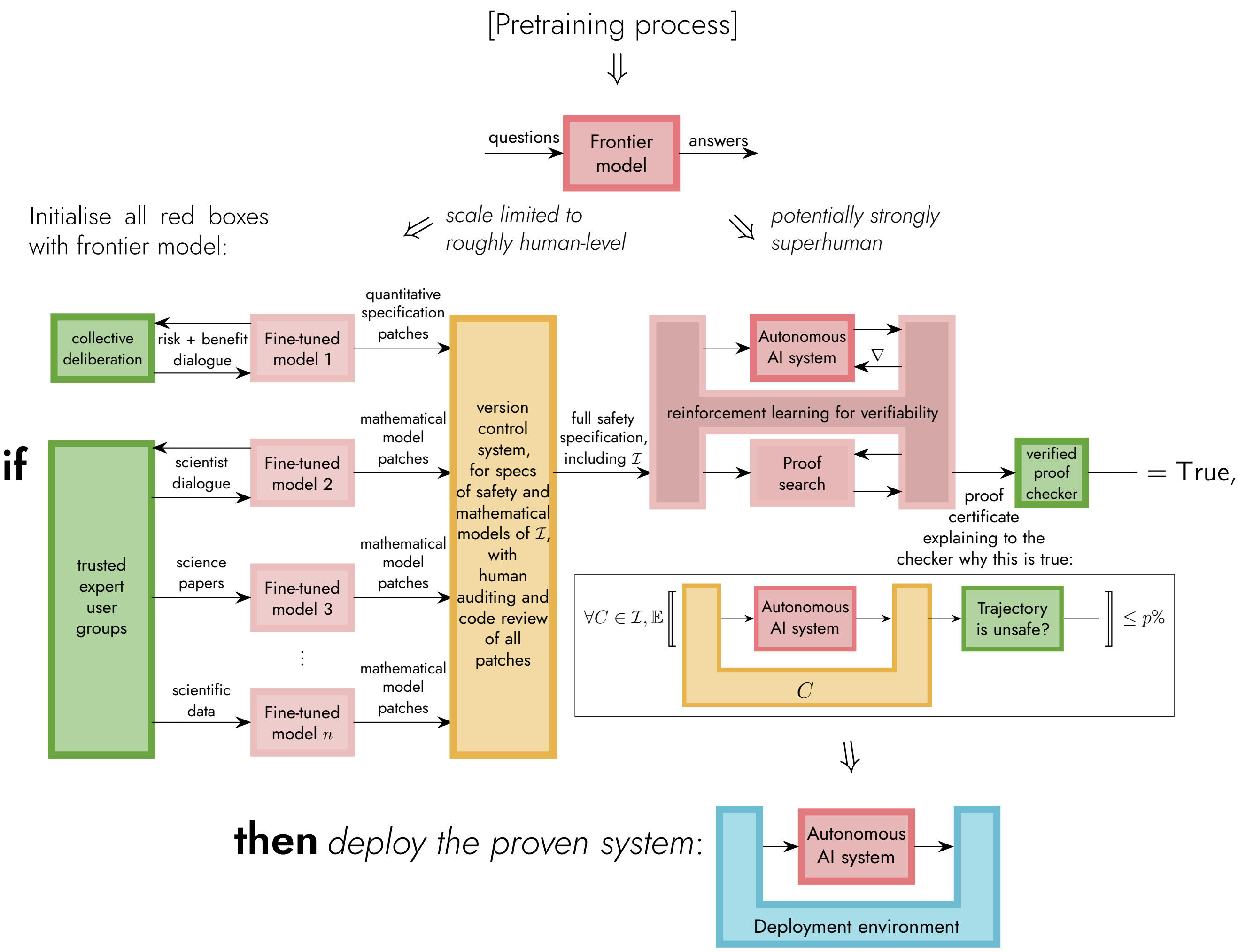
However, very little R&D effort is currently going into approaches that provide quantitative safety guarantees about the deployment—even compared to AI safety as a whole—because this standard is commonly considered either impossible or impracticable: either it won't work, or if it does work, it would take too long and not provide enough value, compared to direct use of frontier models.

One emerging but underexplored approach is the concept of a "gatekeeper" safeguard that formally verifies proof certificates which can be produced by a frontier model itself (with different fine-tuning and scaffolding).

A gatekeeper safeguard would have 3 distinct AI components each building on pre-trained frontier models:

1. frontier models adapted to iteratively construct an explainable, auditable, scientific **mathematical model** of the task-relevant aspects of the real world, and build on this to define **quantitative specifications** of safety criteria (as well as of task success);
2. frontier models adapted to use in-context learning to drive a **proof search** to prove certain probabilistic quantitative bounds (on the behaviour of certain cyber-physical systems when neural networks acting as autonomous AI systems are deployed into them), and produce proof certificates (which can be checked by a proof-checker that is itself formally verified); and
3. a variant of a deep **reinforcement learning training loop** which adapts (by fine-tuning, policy-gradient optimisation, pruning, distillation, low-rank decomposition, or otherwise) a powerful frontier AI model to *become* a neural network with **high verifiability** (to be verified by the previous method).

Figure 1: The "gatekeeper" approach can be seen as an overall workflow to produce safe AI systems.



The goal of this programme is to demonstrate that a "gatekeeper" workflow like this can be a viable, universal solution for safeguarding many economically and socially valuable applications of AI.

We'll do that by demonstrating the following:

TA1. That an extendable, interoperable language and platform can be built to maintain real-world models and specifications and check proof certificates.

TA2. That we can use frontier AI to help domain experts build best-in-class mathematical models of real-world complex dynamics with relevance to valuable applications, and leverage frontier AI to train autonomous systems that can be verified with reference to those models.

TA3. That a gatekeeper-safeguarded autonomous AI system deployed in a critical cyber-physical operating context can unlock significant economic value with quantitative safety guarantees.

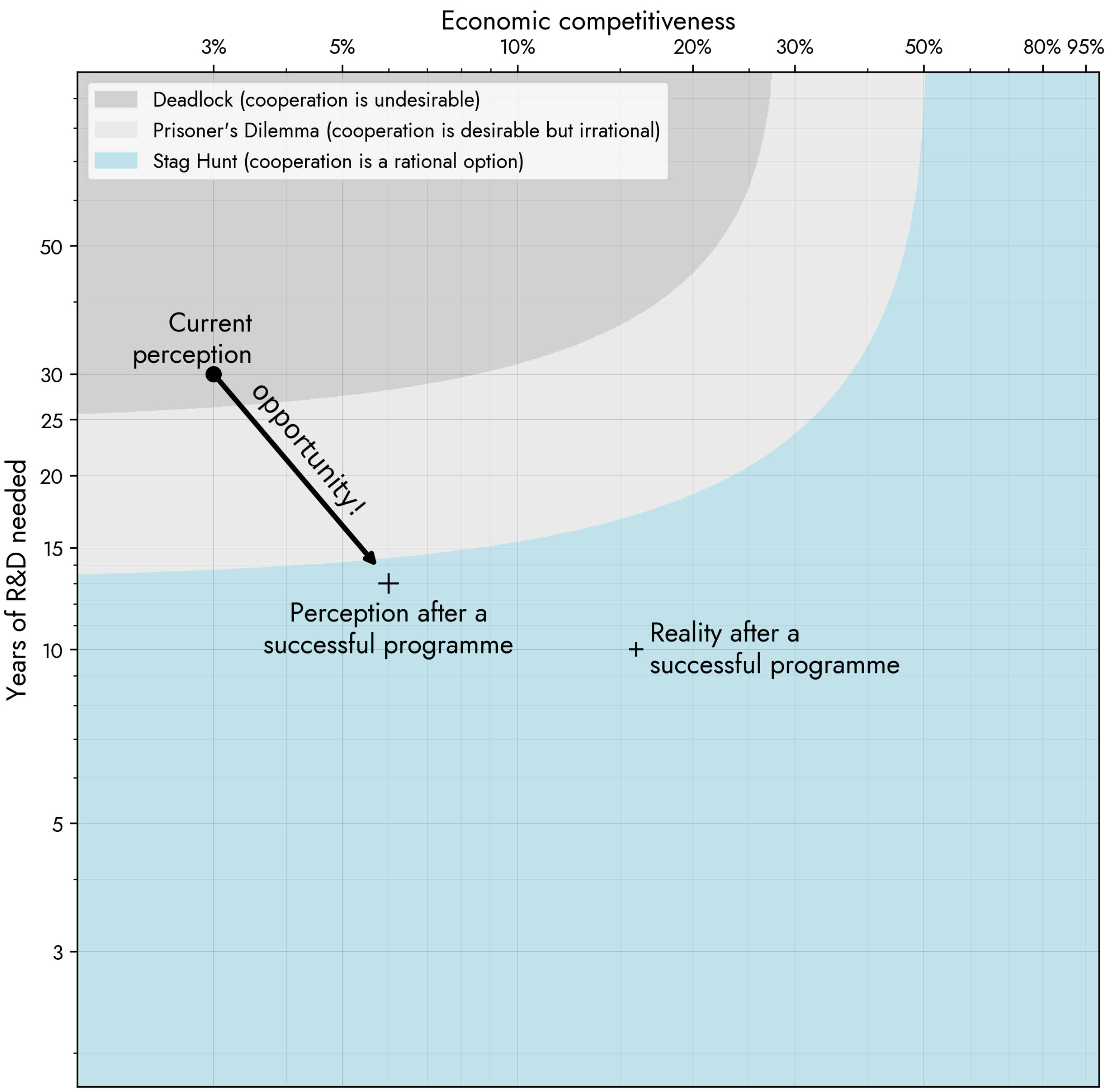
If we're successful with any of these demonstrations, we believe this programme will be valuable. If we succeed in all three, we believe we will have elucidated a viable and scalable path to safe transformative AI.

Our most aspirational theory of change is: if we can **catalyse a global scientific consensus** that

+ there is a feasible R&D pathway which uses frontier AI systems *only* via assemblages that provide quantitative safety guarantees,
  + that one eventual application of these safety-critical assemblages is defending humanity against potential future rogue AIs[5,22] enough to reduce the risks to an acceptable level,
+ and that this milestone could be achieved, thereby making it safe to unleash the full potential of superhuman AI agents, **within a time frame that is short enough** (<15 years),
+ and **with enough economic dividends along the way** (>5% of unconstrained AI's potential value)—

then this would enable a new, cooperative Nash equilibrium in the global strategic landscape around frontier AI, enabling players to agree to follow this pathway (according to some plausible modelling assumptions; see Figure 2 and Appendix B).

Figure 2: In a simplified game-theoretic model of the choice that strategic players face (about whether to commit to a lengthy process of ending the acute risk period while using superhuman AI only inside systems with quantitative safety guarantees), the Nash equilibrium structure depends critically upon the required duration and the intermediate economic benefits (as a fraction of the net profit of unconstrained AI). See Appendix B for more details on these modelling assumptions.



What we expect to fund

The programme is broken out into 3 technical areas (TAs), with the following names and top-level goals:

TA1 **Scaffolding**: Challenging the claim that "it's not possible to formally specify what it means for a system to be safe in the real world" by demonstrating a tool that non-mathematician domain experts can use to develop and refine quantitative models and specifications about their systems of interest (e.g. power grids, epidemics, R&D roadmaps), with cross-domain interoperability (e.g. a lockdown changes the demand patterns on the power grid; a vaccine R&D process changes the epidemic). TA1 also includes a proof language and checker intended for AI systems to hook into to check if their outputs are verifiable (and iterate until they are).

TA2 **Machine Learning**: Challenging the claim that "even if it were possible to specify real-world safety, it wouldn't be economically competitive to train an AI system that *provably* satisfies such a spec" by demonstrating empirically (albeit in simulation) a learned controller for a complex system of significance, that has verifiable quantitative probabilistic safety bounds, and achieves *both* better performance and better resilience to adverse events than non-certified baselines.

TA3 **Applications**: Challenging the claim that "even if you could train AI systems that provably satisfy their specs, no one would really use them, as the economy would only adopt mainstream AI instead" by demonstrating that the barriers to adoption in significantly valuable application domains can be overcome, leading to an organisation maintaining an actual production deployment in practice.

Technical Area 1 (TA1): Scaffolding

The primary requirement for a gatekeeper AI is to prove that an autonomous AI system satisfies its specification, such as a quantitative upper bound on the probability of significantly unsafe consequences of its actions.

In order to make such a proof, one must first define the specification with regard to the AI system; in order to define a specification of safety for an AI system operating as part of a real-world cyber-physical system, one must define a mathematical model of the dynamics of the environment and context into which the system is deployed. The specification can then make demands about what occurs in the environment (e.g. that some formally defined kind of "harm" does not take place with high probability), rather than formal specifications referring only to the relationship between inputs and outputs of the AI system itself (which is sufficient for defining some nontrivial properties, like "adversarial robustness", but not any physical kind of safety). In order to be taken as "ground truth" about the bounds of what might occur in the deployment environment, to serve as a root of trust for the system's certification, these mathematical models must be audited by teams of humans, and therefore the modelling language in which they are expressed must be both human-understandable and amenable to formal methods.

But this language cannot be hard-coded by humans, as the pioneers of good old-fashioned AI imagined. As Sutton put it in The Bitter Lesson[49], "simple ways to think about...the arbitrary, intrinsically-complex, outside world...are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity." As such, the goal of TA1 is not to develop an ontology (e.g. a list of the kinds of entities in the world and their possible relationships), but as much as possible, to develop a *meta*-ontology (e.g. the Semantic Web conceptual framework is a meta-ontology, and the framework of ordinary differential equations is a different meta-ontology) in which interoperable, compositional mathematical artefacts can be developed regarding all the causal pathways by which task-specific AI deployments may cause harm. We envision that these artefacts would be co-developed by human-level or near-human-level AI "copilots", under the supervision of human domain experts, mediated by specialised human-computer interface paradigms designed for this purpose.

Mathematical models are the most fundamental type of artefact in TA1. A mathematical model has rigorous formal semantics that define a state space and a dynamics. At this level of foundations, we would like to transcend assumptions that the state space be finite, discrete, or even finite-dimensional; rather, we would like to constrain the state space only to be σ-compact (a countable union of compact spaces). For cross-domain interoperability, we would like our modelling language to be a common generalisation of many existing modelling languages (see Appendix A.1); and we would like our mathematical models to be constructed within a compositional "doctrine of dynamical systems" with flexible *composition patterns*[37]. In the same spirit of interoperability between modelling frameworks, we would like to transcend assumptions that the dynamics be deterministic, stochastic, nondeterministic, graphical, or temporal; rather, we would like the basic concept of dynamics to be any rule that enables one to deduce information about some observables of the system trajectory from information about others, using an epistemic framework that encompasses both probabilistic (Bayesian) uncertainty and nondeterministic (Knightian) uncertainty within one monad[1].

Formal specifications are predicates about counterfactual probabilities about the distribution of trajectories in the state space. For example, a specification might require an upper-bound on the counterfactual probability of someone being harmed by an AI controller (under a resolution of unknowns in which they would not have been harmed if the AI did nothing):

$$\mathbb{E}_{\omega\sim\Omega}\Big(\exists_{v:V}\mathsf{Harmed}(v)(\mathsf{WorldModel}(S\mapsto\mathsf{AIController})(\omega)) \\ \wedge\neg\mathsf{Harmed}(v)(\mathsf{WorldModel}(S\mapsto\mathsf{DoNothing})(\omega))\Big) < 10^{-4}$$

The definition of state-space predicates such as 'Harmed' can be expressed in the same modelling language as state-space dynamics, but counterfactual queries and probabilistic bounds are additional logical primitives, which are likely best incorporated via an extended language for specifications.

Proof certificates are a quite broad concept, introduced by[33]: a *certifying algorithm* is defined as one that produces enough metadata about its answer that the answer's correctness can be checked by an algorithm which is so simple that it is easy to understand and to formally verify by hand[2]. We are very interested in certificates, because we would like to rely on black-box advanced AI systems to do the hard work of searching for proofs of our desired properties, yet without compromising confidence that the proofs are correct. In this programme, we are specifically interested in certificates of *behavioural properties* of cyber-physical systems (ranging from simple deterministic functions, to complex stochastic hybrid systems incorporating nuances like nondeterminism and partiality). To be a bit more specific, the properties of interest are typically universally quantified statements claiming that if some precondition A is true about a subsystem's state trajectory $x(t)$ at time t_0, then the probability of some postcondition B being true at time t_1 must be within a certain range:

$$\forall\theta\in\Theta,\quad \mathbb{P}\bigg(B\Big(x\big(t_1(\theta)\big),\theta\Big)\;\bigg|\;A\Big(x\big(t_0(\theta)\big),\theta\Big)\bigg)\in\big[l(\theta),u(\theta)\big]$$

Some examples of kinds of certificates that could be useful in proving such quantitative bounds include barrier certificates[41], reach-avoid supermartingales[56], contraction metrics[50], Alethe certificates[3], LFSC proofs[48], branch-and-bound certificates[8], certificates based on abstract interpretation or bound propagation[6,11], and Noetherian induction proofs[47]. Our goal in this programme's proof certificate language is to unify as many of these approaches as possible[3], to give an AI system maximum flexibility in constructing any sound and valid argument for its quantitative safety bounds.

Neural systems must be expressible in the modelling language, since the specifications we want to check will refer to variables which are to be filled in with neural networks, such as the '$\mathsf{AIController}$' variable in the example specification above. This is no problem for the theoretical semantics of the language because neural networks are semantically just continuous functions. However, it is a consideration for the data structures, algorithms, and interface formats, since neural networks tend to be very large, but typically have stereotyped compressible structure in terms of tensor algebra, of which we would of course want to take advantage. It may be useful to build on the ONNX format for the interchange of neural network architectures and weights[4]. We refer to "neural systems" to encompass a broader class of autonomous systems with neural-network components, but which may also include other algorithms[4].

Programmatically, TA1 shall be divided into three technical subareas: **TA1.1 Theory**, **TA1.2 Backend**, and **TA1.3 Human-computer interface**, each of which covers the entire gamut of $\{\text{models}, \text{specifications}, \text{proofs}\}$, but from different perspectives, as follows.

[1] along the lines of the $\mathcal{C}^{\downarrow}$ monad from [35, Definition 36], the "homogeneous ultracontribution" monad $\square^c$ from [29, sec 1.1], or the "convex powerset of distributions monad" [45]

[2] e.g. using an interactive theorem-proving language such as [36]

[3] We envision using the more logical and inductive approaches to organise an uncountable and potentially unbounded state space by Noetherian induction and case analysis of finite partitions, resulting in a finite set of proof obligations regarding only compact state spaces, each of which can be discharged by some primitive quantitative bounding certificate. Some promising recent work along these lines includes [51, 57].

[4] e.g. a runtime-verification decision module that switches between two neural networks

TA1.1 **Theory** shall research and construct computationally practical mathematical representations and formal semantics for world-models, specifications, proofs, neural systems, and "version control" (incremental updates or patches) thereof.

TA1.2 **Backend** shall develop a professional-grade implementation of the Theory, yielding a distributed version control system for knowledge represented as mathematical world-models, as well as for specifications. The Backend shall also maintain a programmatic interface that can be used by AI-driven machine learning training loops to "check in" neural networks and verify proofs about them, with the backend producing counterexamples or informative error messages for invalid proofs. As a stretch goal, the Backend could also be responsible for "compiling" neural networks into a deployable executable package that has a high assurance of correctly implementing the exact mathematical function which was verified.

TA1.3 **Human-computer interface** shall develop a professional-grade user experience for eliciting formal explainable goals from stakeholders; auditing and editing scientific models; interactively collaborating with AI modelling assistants; reviewing proven guarantees and sample trajectories; red-teaming; developing new safety specifications in light of shortfalls; run-time monitoring of whether the incoming data is consistent with the mathematical model of the environment, especially the propositions claimed about it in a safety proof, to spot potentially safety-relevant anomalies; and any other aspects of the programme that are found to require human-computer interaction.

Technical Area 2 (TA2): Machine Learning

Although the "gatekeeper" concept is intended to primarily build on mainstream pre-trained frontier AI models, it involves forking a frontier model and fine-tuning or "post-training" it in a few different ways in parallel, to assemble a workflow which transforms one final fork of the frontier model into a verifiably safeguarded AI system[5].

An advantage of our approach is that the satisfaction of specifications can be quantitatively verified[6], but a substantial risk from developing a recipe for AI systems that certifiably behave in accordance with arbitrary specifications is that those specifications may not be adequately informed by all affected stakeholders. To mitigate this, we have also included sociotechnical control/oversight processes within the scope of TA2.

Programmatically, TA2 shall be divided into four subareas:

TA2.1 **World-modelling ML** shall fine-tune pre-trained (near-)human-level AI systems to be fluent in **TA1**'s new language of scientific world models, to assist teams of human scientists and engineers in formally describing the operating environment and specifications for any given application(s). This could include fine-tuned models for:

+ extracting structured data about physical parameters from spreadsheets[[26]], CAD drawings[[55]], unstructured chart images[[20]], and other formats;
+ proposing ways to use this data to instantiate and populate a composition of probabilistic models, e.g. [53]
+ applying probabilistic data cleaning techniques[[31]] to propose patches which make the data consistent and plausible
+ extracting data from time-series measurements of a system in action, assisting in Bayesian updating on such data: e.g. by learning to approximate posteriors, by discovering compressed latent representations for data, by learning an amortised approximate likelihood, etc.[[24,42]]
+ engaging in structured dialogue threads regarding model components (somewhat analogous to *code review*[[39]]), with humans or potentially with each other
+ extracting mathematical models from scientific papers, e.g. [14]
+ conjecturing mathematical models from data alone, e.g. [46]

TA2.2 **Proof-search ML** shall fine-tune frontier AI systems as tools that search for proof certificates[7] to certify safety properties of cyber-physical systems with learned components, i.e. that establish an upper bound on the probability of a learned controller violating a safety specification, to be proof-checked by the **TA1.2** Backend, given assumptions from the world-models output by the **TA2.1** World-modelling ML.

[5] through modified RL training, and likely various forms of compression, such as pruning, distillation, or low-rank decomposition

[6] potentially even by mutually distrusting entities, via zero-knowledge proofs (e.g. [28]); these come at a substantial but declining computational cost

[7] of various potential kinds, e.g. [3, 6, 8, 10, 11, 15, 41, 47, 48, 56, 57]; see the paragraph about proof certificates on page 6

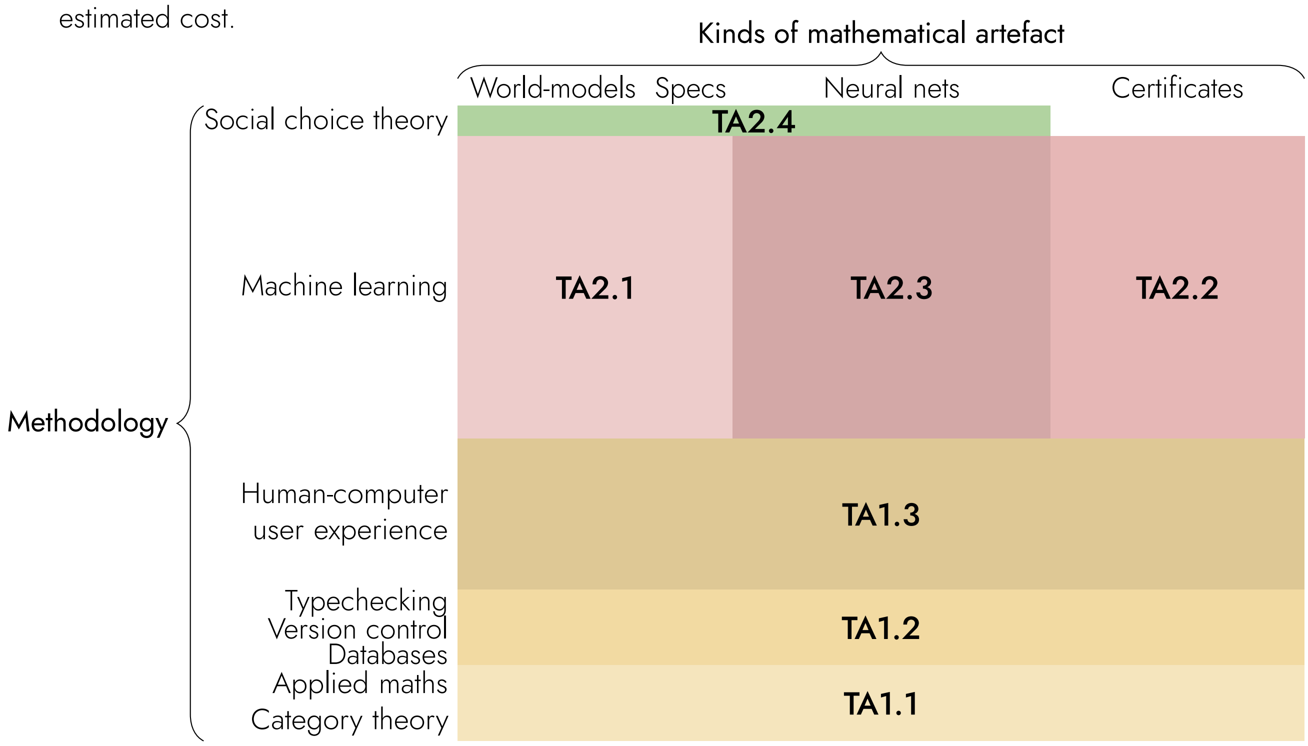
TA2.3 **Training for certifiable ML** shall develop an automated "training loop" which is similar to reinforcement learning, but produces as output an autonomous decision-making system which can be certified as satisfying the specification, by the proof-search of **TA2.2**. Potential pathways to achieve this include:

+ Counterexample-Guided RL for Static Verification (e.g. [27]): If the verifier produces specific examples[8] where the specifications fail, these can be incorporated into a training loop as additional data samples, to augment the usual Monte Carlo rollout trajectories.
+ Runtime Verification (e.g. [34]): Instead of learning a single end-to-end neural network which is certifiable, the autonomous decision-making system which is certified could *contain* a version of a verifier that is fast enough to run in real-time[9], which has different tradeoffs.

  Also, runtime verification is useful even if a single neural network is certified using static verification, because it would enhance overall sociotechnical robustness to use runtime verification techniques as runtime monitoring of functions of the sensor inputs and state estimates that play a role in the safety proof certificate, to proactively identify potential anomalies in which the true, real-world deployment environment diverges from what was modelled in an unexpected and potentially safety-relevant way.
+ Probabilistic Shields (e.g. [25, 54]): Runtime verification can also be incorporated into a system *during training* if the verifier is differentiable. The combination of differentiability, soundness, and sufficient computational speed for real-time use may not be tractable for nontrivial world-models, although a potential breakthrough here would be exciting. Existing shielding techniques may be useful for reward shaping, alongside runtime verification and/or static verification.

TA2.4 **Sociotechnical gate** shall develop and amend processes for diverse groups of stakeholders to make collective deliberations about acceptable risks and safety specifications, suggest quantitative bargaining solutions that could facilitate multi-objective certifiable ML, and make go/no-go decisions about any new deployment, release, or publication. This will use the quantitative safety guarantees computed in **TA2.2**, via the human-computer interfaces produced in **TA1.3**, to assure that benefits exceed risks in expectation.

Figure 3: One way of visualising the topical breakdown of subareas in TA1 and TA2 is by considering kinds of artefact and methodology as orthogonal axes. Note that while most of the machine learning tasks in TA2 can be decoupled, the TA1 scaffoldings will overlap substantially. Size approximates estimated cost.



[8] regions of the state-space Ω of exogenous conditions with a nontrivial probability measure

[9] In the "black-box simplex architecture" [34], the verifier must certify before every action is output that, after that action is taken, a *backup controller* is still very likely capable of stabilising the state into the safe region of state space (without optimising for or achieving anything other than being verifiably safe). If the certification ever fails or runs out of time for a given action, the system switches to the backup controller, which was proven at the previous time-step to be able to recover to a stable safe state.
Note that such a system can be certified by an inductive proof.

Technical Area 3 (TA3): Applications

Ultimately, it does not matter how safe a system is unless it is an acceptable substitute for less safe alternatives. **TA3**'s goal is to demonstrate that "gatekeeper AI" as a workflow can be used to create and maintain decision-support tools and/or safeguarded autonomous AI systems that deliver value in practice for specific tasks. Programmatically, TA3 will likely consist of 2—4 full teams, each pursuing a different application area using the tools developed in the other TAs. An initial **TA3 Phase 1** will cast a wider net, funding a larger number of part-time efforts to elicit requirements in application areas and draft models and specifications by hand, in advance of the earliest prototypes being made available by other TAs.

Application areas suitable for an early demonstration (i.e. within our programme duration) likely fit these criteria:

(a) **scalability**— an ideal application area can offer a family of problems with "instances" at various scales of the size or number of entities being modelled, something like this:
   + with n_1 it is almost trivial,
   + with n_2 it is analytically tractable, but tricky,
   + with n_3 it is already practically interesting, but routine for baseline methods,
   + with n_4 it is pushing the limits of what seems practical today,
   + and if we could make it practical with n_5, that would be a game-changer

(b) **known in principle**— the primary difficulties involved in this problem area should ***not*** include:
   + lack of a solid informal scientific consensus understanding of substantial aspects
   + difficulty of making sufficiently detailed measurements or observations of the phenomenon

   Instead, the difficulties should be more like "there's just a lot of moving parts"
   or (less preferably) "it's just really inefficient to compute"

(c) **predictable in principle**— not swamped by sensitivity to initial conditions

(d) **need for high trust**— because of their safety-critical or mission-critical nature, automation and AI solutions for this application are currently facing serious barriers to adoption due to lack of reliability, which our methods could directly address

(e) **absence of bias** or **bias mitigation strategy**— either we have data which is unaffected by systemic bias, and/or we have no reason to expect existing large language models distilled from internet text to bring in systemic bias, or we have a plan in place to avoid the default outcome in which these biases become encoded and amplified by the model[10]

(f) **large-scale relevance**— if humanity mastered automated control over this phenomenon at the larger scales, it could provide socioeconomic benefits on the scale of hundreds of millions of pounds per year

(g) **existing baseline predictor or controller**(s)— there's some approximate and/or costly ways that large instances are dealt with in practice today, to which new predictive mathematical models and new decision-making systems could be quantitatively compared

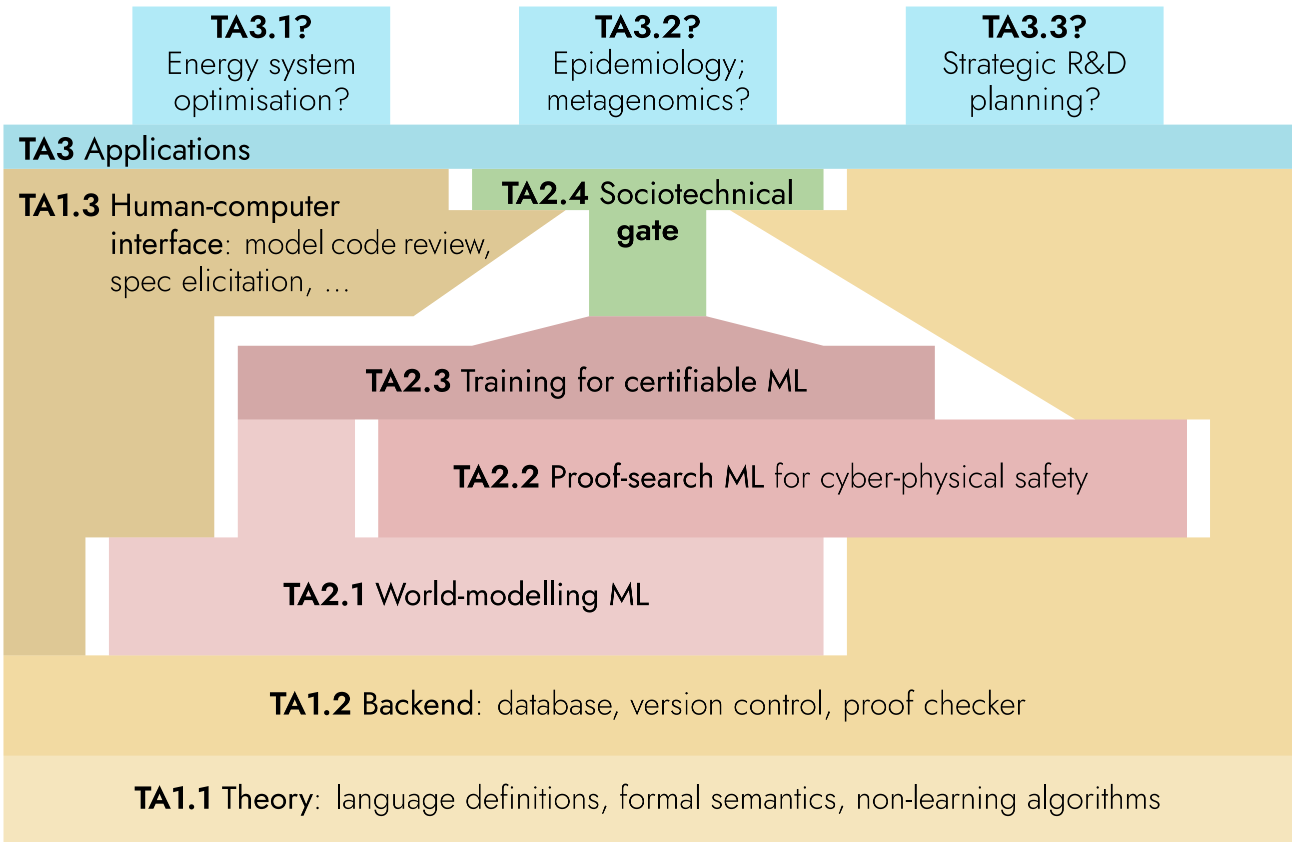
The full set of specific applications is to be determined in the review of responses to the **TA3** solicitation, but some areas we're currently exploring include:

+ **energy system** optimisation, e.g.
   + real-time power dispatch to manage supply and demand and respond to perturbations (especially complex with energy storage and more renewable supply)
   + probabilistic demand forecasting, on various time scales
   + keeping network models up-to-date (e.g. with new rooftop solar installations)
   + long-term planning of transmission network capacity improvements
+ **telecommunications** network optimisation
   + real-time allocation of beamforming subchannels to optimise transmitter energy consumption
   + management of upstream/backhaul capacity
   + long-term network expansion planning

[10] Even with a fully explainable and transparent model, there will be free parameters, and if applied to an area such as the dynamics of crime or social outcomes, both the parameters and the overall structure of the model can encode bias in hard-to-notice ways.

+ **supply chain** and inventory management
    + probabilistic demand forecasting
    + distribution requirements planning
    + last-mile delivery routing
+ control systems for **robots in human environments**
+ **medical device** control systems
+ optimisation of **clinical trials**
+ infectious disease **epidemiology**
    + especially under various intervention scenarios, for decision support
    + incorporating diverse data sources, including metagenomics
+ **climate** and weather prediction
    + especially under various intervention scenarios
+ **transport** optimisation
+ aircraft and spacecraft **flight dynamics**
    + fully autonomous
    + autopilots
    + airspace/traffic control
+ **R&D planning**
    + roadmapping
    + short-term project management
    + medium-term forecasting
    + long-term R&D portfolio planning
+ complex **business processes**
+ **data integration** in contexts where it is usually done by hand to avoid mistakes

Figure 4: The **interfaces** between all technical subareas can be shown visually as horizontal contacts.



How we expect to fund

We anticipate staging funding opportunities in the following sequence:

TA1.1 **Theory**

In this area we would fund researchers for one or two projects each that would be the initial hypothesis for what they would start working on, but we would hold this hypothesis lightly and assume that on a quarterly or even monthly cadence, it might make sense to change course and take on a different problem—or, having definitively solved a certain scoped question, then build on that solution to identify the shape of the next frontier.

We would suggest an initial list of problems (see Appendix A for an early draft), and would welcome proposals to tackle those problems head-on, but we are also open to suggestions of related but distinct theoretical problems.

We expect to have an ongoing collaboration and free flow of ideas between participants in **TA1.1**. In this area we would in large part evaluate success by how much participants have built on others' work and how much others have built on their work, and in part by subjective review.

TA2 **Machine learning** (Phase 0)

In this area we would plan to fund a major R&D effort within a single institution, ideally with the following characteristics:

+ Based in the United Kingdom
+ Co-funded by one or more partner organisations
+ World-class cybersecurity
+ Credible ability to source world-class talent in machine learning research & engineering
+ Decisions to publish algorithms, models, or code, or to release products or API access externally, should be governed by a diverse board with the sole mission of ensuring that the expected benefits of AI to humanity and society substantially exceed the risks
+ Flexibility to pursue multilateral information-sharing and strategic partnerships with other private and/or government-sponsored entities— if and only if determined to align with the mission

Such an institution could be a unit or subsidiary of an existing organisation, or it could be a newly formed entity. An early Phase 0 would fund initial explorations to put together a full proposal.

Success in TA2 would be evaluated by one or more groups in **TA3** **Applications**, which would each be building benchmark metrics for performance in a specific application area, such as energy system optimisation, autonomous aircraft, R&D planning, dexterous manipulation, etc., with one area being selected for the initial scope of work.

TA1.2 **Backend**

In this area we would fund 1 or 2 software development organisations (with strong mathematics capabilities) to elicit concrete requirements from **TA1.1** creators for implementation of their theory. If the requirements engineering process is successful, this would lead to a much larger award to build some or all of the backend software for the programme's software platform (with success being evaluated according to those requirements).

TA1.3 **Human-computer interface**

In this area we would fund 1 or 2 software development organisations (with strong design/HCI/UX capabilities) to begin a collaborative process of shaping the requirements for interfaces that can help humans with diverse ways of thinking to interact with the systems being built in **TA1.2** **Backend** and **TA2.1** **World-modelling ML**. In this area, success would be evaluated by reviews from users across all areas of the programme.

TA3 **Applications**

In this area we would solicit potential entrepreneurs or existing entities interested in using our gatekeeper AI workflow to build safeguarded products for specific tasks in a specific sector, with Phase 1 providing a small amount of funding to deeply understand customer needs and elicit requirements, begin to source datasets, design evaluation suites to validate the performance of predictive models and autonomous or semi-autonomous controllers, etc.

We anticipate that this will be a highly coordinated programme, with quarterly workshops to facilitate teams with interfaces (the horizontal contact surfaces shown in Figure 4) having opportunities to reach agreements about syntax and semantics of the formats of information that would flow through such interfaces.

In advance of programme launch we will coordinate with the UK's AI Safety Institute (AISI) to identify any potential areas of collaboration.

Intellectual property will be managed differently in each TA:

TA1 work is to be carried out in public by default, with permissively licensed open-source code and documentation, no patents without a patent non-aggression pledge (example), and all publications available open-access.

This is primarily to accelerate adoption and flow of ideas, but also because in the ultimate vision, the TA1 scaffolding is the platform for a global assurance mechanism that enables multiple actors to verify certificates from each other's AI systems proving compliance with internationally agreed norms; the involvement of a patchwork of proprietary IP rights would complicate such usage.

TA2 work is to be done in a secure environment, with serious measures in place to avoid leaks of model weights (or even leaks of most concrete algorithmic ideas), for example to include strict NDAs, device policies, etc. Patents may be filed without a patent non-aggression pledge if the TA2 entity sees fit, but most patentable inventions in TA2 should more likely be protected as trade secrets.

The TA2 entity should have a robust process in place to review and wisely evaluate potentially beneficial releases (publications, weights, API availability, commercial licensing—including to TA3 entities, etc.).

This is because, if successful, TA2 work would substantially facilitate AI misuse in addition to reducing the risk of AI accidents[5], so the outputs must be carefully governed to ensure a net-positive impact, which implies that in the first instance they must not proliferate irreversibly.

TA3 work, which consists of vertical-domain-specific models, libraries, techniques, and control systems constructed by TA3 creators using TA1 and TA2 software tools, can be treated in the ordinary way as the proprietary IP of its creators.

What we are still trying to figure out

+ What are the initial 2–4 target application areas and tasks where **TA3** will aim to demonstrate new safe capabilities?
+ What are the shortest critical paths for each subarea to get started, e.g. by using a preliminary version of each dependency, or by acting as an observer?
+ In what timeline can we sequence solicitations and kickoffs?
+ What is the best format for high-bandwidth interactions to elicit requirements and establish interfaces and abstractions?
+ Will we attract strong enough participants within the UK—or *to* the UK—for **TA2**?

ENGAGE

Our next step is to launch a funding opportunity derived or adapted from this programme thesis.

Click **here** to register your interest, or to provide feedback that can help improve our thinking.

SOURCES

References cited in this document.

[1] **Alvarez-Picallo, Mario**, Dan R. **Ghica**, David **Sprunger** and Fabio **Zanasi**. Functorial string diagrams for reverse-mode automatic differentiation, 28th July 2021, arXiv: `2107.13433` (cited on p. 16)

[2] **Amadini, Roberto**, Graeme **Gange**, Peter **Schachte**, Harald **Søndergaard** and Peter J. **Stuckey**. Abstract interpretation, symbolic execution and constraints, Schloss Dagstuhl, 2020, DOI: `10.4230/OASIcs.Gabbrielli.7` (cited on p. 17)

[3] **Andreotti, Bruno**, Hanna **Lachnitt** and Haniel **Barbosa**. Carcara: An efficient proof checker and elaborator for SMT proofs in the Alethe format, *TACAS 2023*, vol. 13393, LNCS, Springer, 2023, 367–86, DOI: `10.1007/978-3-031-30823-9_19` (cited on pp. 6, 7)

[4] **Bai, Junjie**, Fang **Lu**, Ke **Zhang et al.** ONNX: Open neural network exchange, 2019, URL: `https://github.com/onnx/onnx` (cited on p. 6)

[5] **Bengio, Yoshua**. AI and catastrophic risk, *Journal of Democracy* 34.4 (Oct. 2023), 111–21, DOI: `10.1353/jod.2023.a907692`, URL: `https://www.journalofdemocracy.org/ai-and-catastrophic-risk/` (cited on pp. 1, 4, 12)

[6] **Besson, Frédéric**, Thomas **Jensen** and David **Pichardie**. A PCC architecture based on certified abstract interpretation, Research Report RR-5751, Inria, 2005, URL: `https://inria.hal.science/inria-00070268` (cited on pp. 6, 7)

[7] **Boisseau, Guillaume**, Chad **Nester** and Mario **Román**. Cornering optics, *EPTCS 2023*, vol. 380, 6, 2023, 97–110, DOI: `10.4204/EPTCS.380.6`, arXiv: `2205.00842` (cited on p. 16)

[8] **Cheung, Kevin K. H.**, Ambros **Gleixner** and Daniel E. **Steffy**. Verifying integer programming results, *IPCO 2017*, vol. 10328, LNCS, Springer, 2017, 148–60, DOI: `10.1007/978-3-319-59250-3_13`, arXiv: `1611.08832` (cited on pp. 6, 7)

[9] **Dash, Swaraj**, Younesse **Kaddar**, Hugo **Paquet** and Sam **Staton**. Affine monads and lazy structures for Bayesian programming, *POPL 2023*, vol. 7, ACM, Jan. 2023, 1338–68, DOI: `10.1145/3571239`, arXiv: `2212.07250` (cited on p. 16)

[10] **Dawson, Charles**, Sicun **Gao** and Chuchu **Fan**. Safe control with learned certificates: A survey of neural Lyapunov, barrier, and contraction methods for robotics and control, *IEEE Transactions on Robotics* 39.3 (June 2023), 1749–67, DOI: `10.1109/tro.2022.3232542`, arXiv: `2202.11762` (cited on p. 7)

[11] **Desmartin, Remi**, Omri **Isac**, Grant **Passmore**, Kathrin **Stark**, Ekaterina **Komendantskaya** and Guy **Katz**. Towards a certified proof checker for deep neural network verification, *LOPSTR 2023*, vol. 14330, LNCS, Springer, 16th Oct. 2023, 198–209, DOI: `10.1007/978-3-031-45784-5_13`, arXiv: `2307.06299` (cited on pp. 6, 7)

[12] **Di Lavore, Elena** and Mario **Román**. Evidential decision theory via partial Markov categories, *LICS 2023*, IEEE, 30th Jan. 2023, arXiv: `2301.12989` (cited on p. 17)

[13] **Dokter, J. L.** Analysis methods of hybrid systems applied to stochastically and dynamically colored Petri nets, Master's Thesis in Applied Mathematics, University of Groningen, 10th May 2013, URL: `https://fse.studenttheses.ub.rug.nl/10947/1/Scriptie.pdf` (cited on p. 16)

[14] **Dunn, Alexander**, John **Dagdelen**, Nicholas **Walker**, Sanghoon **Lee**, Andrew S. **Rosen et al.** Structured information extraction from complex scientific text with fine-tuned large language models, 10th Dec. 2022, arXiv: `2212.05238` (cited on p. 7)

[15] **Edwards, Alec**, Andrea **Peruffo** and Alessandro **Abate**. Fossil 2.0: Formal certificate synthesis for the verification and control of dynamical models, 16th Nov. 2023, arXiv: `2311.09793` (cited on p. 7)

[16] **Everidj, M. H. C.** and H. A. P. **Blom**. Hybrid state Petri nets which have the analysis power of stochastic hybrid systems and the formal verification power of automata, tech. rep. NLR-TP-2010-324, NLR Air Transport Safety Institute, Feb. 2010, URL: `https://reports.nlr.nl/server/api/core/bitstreams/167f434f-8a4a-4cb1-8a6f-17e41e72c659/content` (cited on p. 16)

[17] **Grandis, Marco**. Higher dimensional categories: From double to multiple categories, World Scientific, Apr. 2019, DOI: `10.1142/11406` (cited on p. 16)

[18] **Gruetzemacher, Ross** and Jess **Whittlestone**. The transformative potential of artificial intelligence, *Futures* 135 (Jan. 2022), 102884, DOI: `10.1016/j.futures.2021.102884` (cited on p. 2)

[19] **Hadzihasanovic, Amar** and Diana **Kessler**. Higher-dimensional subdiagram matching, 18th Apr. 2023, arXiv: `2304.09216` (cited on p. 16)

[20] **Han, Yucheng**, Chi **Zhang**, Xin **Chen**, Xu **Yang**, Zhibin **Wang et al.** ChartLlama: A multimodal LLM for chart understanding and generation, 2023, arXiv: `2311.16483` (cited on p. 7)

[21] **Hendrycks, Dan**, Geoffrey **Hinton**, Yoshua **Bengio**, Demis **Hassabis**, Sam **Altman et al.** Statement on AI risk, 2023, URL: `https://www.safe.ai/statement-on-ai-risk` (cited on p. 2)

[22] **Hendrycks, Dan**, Mantas **Mazeika** and Thomas **Woodside**. An overview of catastrophic AI risks, 21st June 2023, arXiv: 2306.12001 (cited on pp. 1, 4)

[23] **Hubinger, Evan**, Carson **Denison**, Jesse **Mu**, Mike **Lambert**, Meg **Tong et al.** Sleeper agents: Training deceptive LLMs that persist through safety training, 10th Jan. 2024, arXiv: 2401.05566 (cited on p. 2)

[24] **Jain, Moksh**, Tristan **Deleu**, Jason **Hartford**, Cheng-Hao **Liu**, Alex **Hernandez-Garcia** and Yoshua **Bengio**. GFlowNets for AI-driven scientific discovery, *Digital Discovery* 2 (5th Apr. 2023), 557–77, DOI: 10.1039/D3DD00002H, arXiv: 2302.00615 (cited on p. 7)

[25] **Jansen, Nils**, Bettina **Könighofer**, Sebastian **Junges**, Alexandru C. **Serban** and Roderick **Bloem**. Safe reinforcement learning via probabilistic shields, 16th July 2018, arXiv: 1807.06096 (cited on p. 8)

[26] **Jia, Ran**, Qiyu **Li**, Zihan **Xu**, Xiaoyuan **Jin**, Lun **Du et al.** SheetPT: Spreadsheet pre-training based on hierarchical attention network, *AAAI* 37.11 (June 2023), 12951–8, DOI: 10.1609/aaai.v37i11.26522 (cited on p. 7)

[27] **Jin, Peng**, Jiaxu **Tian**, Dapeng **Zhi**, Xuejun **Wen** and Min **Zhang**. Trainify: A CEGAR-driven training and verification framework for safe deep reinforcement learning, *CAV 2022*, vol. 13371, LNCS, Springer, 2022, 193–218, DOI: 10.1007/978-3-031-13185-1_10 (cited on p. 8)

[28] **Kitagawa, Fuyuki**, Takahiro **Matsuda** and Takashi **Yamakawa**. NIZK from SNARG, *TCC 2020*, vol. 12550, LNCS, Springer, 2020, 567–95, DOI: 10.1007/978-3-030-64375-1_20 (cited on p. 7)

[29] **Kosoy, Vanessa** and Alexander **Appel**. Infra-Bayesian physicalism: A formal theory of naturalized induction, 1st Dec. 2021, URL: https://www.alignmentforum.org/posts/gHgs2e2J5azvGFatb/infra-bayesian-physicalism-a-formal-theory-of-naturalized (cited on pp. 6, 16)

[30] **Landin, P. J.** The next 700 programming languages, *Communications of the ACM* 9.3 (Mar. 1966), 157–66, ISSN: 1557-7317, DOI: 10.1145/365230.365257, URL: https://www.cs.cmu.edu/~crary/819-f09/Landin66.pdf (cited on p. 16)

[31] **Lew, Alexander K.**, Monica **Agrawal**, David **Sontag** and Vikash K. **Mansinghka**. PClean: Bayesian data cleaning at scale with domain-specific probabilistic programming, (2020), arXiv: 2007.11838 (cited on p. 7)

[32] **Master, Jade**. Petri nets based on Lawvere theories, *Mathematical Structures in Computer Science* 30.7 (Aug. 2020), 833–64, DOI: 10.1017/s0960129520000262 (cited on p. 16)

[33] **McConnell, R.M.**, K. **Mehlhorn**, S. **Näher** and P. **Schweitzer**. Certifying algorithms, *Computer Science Review* 5.2 (May 2011), 119–61, DOI: 10.1016/j.cosrev.2010.09.009, URL: http://alg.cs.uni-kl.de/en/team/schweitzer/publikationen/docs/CertifyingAlgorithms.pdf (cited on p. 6)

[34] **Mehmood, Usama**, Sanaz **Sheikhi**, Stanley **Bak**, Scott A. **Smolka** and Scott D. **Stoller**. The black-box simplex architecture for runtime assurance of autonomous CPS, *NASA Formal Methods Symposium 2022*, vol. 13260, LNCS, Springer, 2022, 231–50, DOI: 10.1007/978-3-031-06773-0_12, arXiv: 2102.12981 (cited on p. 8)

[35] **Mio, Matteo**, Ralph **Sarkis** and Valeria **Vignudelli**. Combining nondeterminism, probability, and termination: Equational and metric reasoning, *LICS 2021*, IEEE, June 2021, DOI: 10.1109/lics52264.2021.9470717, arXiv: 2012.00382 (cited on pp. 6, 16)

[36] **Moura, Leonardo de** and Sebastian **Ullrich**. The Lean 4 theorem prover and programming language, *CADE 2021*, vol. 12699, LNCS, Springer, 2021, 625–35, DOI: 10.1007/978-3-030-79876-5_37 (cited on p. 6)

[37] **Myers, David Jaz**. Categorical systems theory, 3rd Sept. 2023, URL: http://davidjaz.com/Papers/DynamicalBook.pdf, book draft (cited on pp. 5, 16)

[38] **Myers, David Jaz**. String diagrams for double categories and equipments, 2016, arXiv: 1612.02762 (cited on p. 16)

[39] **Pascarella, Luca**, Davide **Spadini**, Fabio **Palomba**, Magiel **Bruntink** and Alberto **Bacchelli**. Information needs in contemporary code review, *Proceedings of the ACM on human-computer interaction* 2.CSCW (Nov. 2018), 1–27, DOI: 10.1145/3274404 (cited on p. 7)

[40] **Patterson, Evan**, Owen **Lynch** and James **Fairbanks**. Categorical data structures for technical computing, *Compositionality* 4.5 (2022), DOI: 10.32408/compositionality-4-5, arXiv: 2106.04703 (cited on p. 16)

[41] **Prajna, Stephen**, Ali **Jadbabaie** and George J. **Pappas**. A framework for worst-case and stochastic safety verification using barrier certificates, *IEEE Transactions on Automatic Control* 52.8 (Aug. 2007), 1415–28, DOI: 10.1109/tac.2007.902736 (cited on pp. 6, 7)

[42] **Radev, Stefan T**, Marvin **Schmitt**, Lukas **Schumacher**, Lasse **Elsemüller**, Valentin **Pratz et al.** Bayesflow: Amortized Bayesian workflows with neural networks, 2023, arXiv: 2306.16015 (cited on p. 7)

[43] **Reutter, David** and Jamie **Vicary**. High-level methods for homotopy construction in associative n-categories, 11th Feb. 2019, arXiv: 1902.03831 (cited on p. 16)

[44] **Schröer, Philipp**, Kevin **Batz**, Benjamin Lucien **Kaminski**, Joost-Pieter **Katoen** and Christoph **Matheja**. A deductive verification infrastructure for probabilistic programs, *OOPSLA 2023*, ACM, Oct. 2023, DOI: `10.1145/3622870`, arXiv: `2309.07781` (cited on p. 17)

[45] **Staton, Sam** and Urs **Schreiber**. Convex powerset of distributions monad, nLab, 4th Dec. 2023, URL: `https://ncatlab.org/nlab/show/convex+powerset+of+distributions+monad` (cited on p. 6)

[46] **Stephany, Robert** and Christopher **Earls**. PDE-READ: Human-readable partial differential equation discovery using deep learning, *Neural Networks* 154 (Oct. 2022), 360–82, DOI: `10.1016/j.neunet.2022.07.008` (cited on p. 7)

[47] **Stratulat, Sorin**. Mechanically certifying formula-based Noetherian induction reasoning, *Journal of Symbolic Computation* 80 (May 2017), 209–49, DOI: `10.1016/j.jsc.2016.07.014` (cited on pp. 6, 7)

[48] **Stump, Aaron**, Duckki **Oe**, Andrew **Reynolds**, Liana **Hadarean** and Cesare **Tinelli**. SMT proof checking using a logical framework, *Formal Methods in System Design* 42.1 (July 2012), 91–118, DOI: `10.1007/s10703-012-0163-3`, URL: `http://homepage.divms.uiowa.edu/~ajreynol/fmsd12.pdf` (cited on pp. 6, 7)

[49] **Sutton, Richard**. The Bitter Lesson, 13th Mar. 2019, URL: `http://www.incompleteideas.net/IncIdeas/BitterLesson.html` (cited on p. 5)

[50] **Tsukamoto, Hiroyasu**, Soon-Jo **Chung** and Jean-Jacques E. **Slotine**. Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview, *Annual Reviews in Control* 52 (2021), 135–69, DOI: `10.1016/j.arcontrol.2021.10.001`, arXiv: `2110.00675` (cited on p. 6)

[51] **Wang, Xinyu**, Luzia **Knoedler**, Frederik Baymler **Mathiesen** and Javier **Alonso-Mora**. Simultaneous synthesis and verification of neural control barrier functions through branch-and-bound verification-in-the-loop training, 2023, arXiv: `2311.10438` (cited on p. 6)

[52] **Weitzman, Martin L.** Gamma discounting, *American Economic Review* 91.1 (2001), 260–71 (cited on p. 19)

[53] **Wong, Lionel**, Gabriel **Grand**, Alexander K. **Lew**, Noah D. **Goodman**, Vikash K. **Mansinghka et al.** From word models to world models: Translating from natural language to the probabilistic language of thought, 2023, arXiv: `2306.12672` (cited on p. 7)

[54] **Yang, Wen-Chi**, Giuseppe **Marra**, Gavin **Rens** and Luc De **Raedt**. Safe reinforcement learning via probabilistic logic shields, 6th Mar. 2023, arXiv: `2303.03226` (cited on p. 8)

[55] **Zhang, Shuming**, Zhidong **Guan**, Hao **Jiang**, Tao **Ning**, Xiaodong **Wang** and Pingan **Tan**. Brep2seq: A dataset and hierarchical deep learning network for reconstruction and generation of computer-aided design models, *Journal of Computational Design and Engineering* (19th Jan. 2024), DOI: `10.1093/jcde/qwae005` (cited on p. 7)

[56] **Žikelić, Đorđe**, Mathias **Lechner**, Thomas A. **Henzinger** and Krishnendu **Chatterjee**. Learning control policies for stochastic systems with reach-avoid guarantees, *AAAI 2023* 37.10 (June 2023), 11926–35, DOI: `10.1609/aaai.v37i10.26407` (cited on pp. 6, 7)

[57] **Žikelić, Đorđe**, Mathias **Lechner**, Abhinav **Verma**, Krishnendu **Chatterjee** and Thomas A. **Henzinger**. Compositional policy learning in stochastic control systems with formal guarantees, 2023, arXiv: `2312.01456` (cited on pp. 6, 7)

[58] **Zilberstein, Noam**, Derek **Dreyer** and Alexandra **Silva**. Outcome logic: A unifying foundation for correctness and incorrectness reasoning, *OOPSLA 2023*, ACM, Oct. 2023, DOI: `10.1145/3586045`, URL: `https://www.cs.cornell.edu/~noamz/files/pubs/outcome.pdf` (cited on p. 17)

Appendix A Early draft of example questions to be answered in TA1.1

The following questions were developed in advance of this programme's first workshop in December 2023.

1. **Compositional knowledge representation**: Is there a natural framework, along the lines of ACSets[40], which incorporates schemas as "first-class citizens", combines the power of relational and algebraic data types, and facilitates a generic notion of version control for changes to types and schemas (alongside incremental computation over data)?

2. **Unified diagram languages**: What are the relationships between ACSets, multiple categories (in the sense of[17]), and various notions of string diagram (e.g. Zanasi et al's hierarchical hypernets[1], Hadzihasanovic's higher-dimensional rewriting[19], Vicary's associative n-categories[43], Myers' string diagrams for double categories and equipments[38], Boisseau et al's cornering diagrams[7], Master's generalised Petri nets[32], etc.)?

   + In particular, can we build a natural "big tent" framework along the lines of "The Next 700 Programming Languages"[30] but for categorical systems diagram languages?

3. **Doctrines of stochastic hybrid systems**: What is the natural construction of a dynamical systems doctrine (in the sense of[37]) for open SDCPNs (stochastic dynamic coloured Petri nets)[13], or equivalently (up to bisimilarity), GSHSs (general stochastic hybrid systems)[16]?

   + As a step in that direction, what about "open jump-drift-diffusion equations"?
   + Can we extend this to PDEs and SPDEs?

4. **Quantitative bounds**: Can we build a "proof system" for establishing convex bounds on the probability distributions of certain variables in a hybrid system, where the proof can invoke abstractions and approximations that have certified error bounds?

   + Can we apply this to basis-theoretic discretisations of PDE solutions such as those used in numerical implementations?

5. **Epistemic conservatism**: Probabilism is more conservative than determinism, but nondeterminism (sometimes called "possibilism") and partiality (or "nontermination") are distinct forms of epistemic conservatism that are not dominated by probabilism. The essential reason we cannot leave out nondeterminism is that not all spaces of possibilities that need to be considered come equipped with a probability measure, or even a canonical base measure like Lebesgue or Haar (with respect to which a uniform probability measure could be defined).

   + [35, Definition 36] introduced a monad that combines these three algebraic effects in a natural way [35, Theorem 38]. Much the same structure (non-empty topologically-closed $\perp$-closed convex sets of subprobability distributions) was discovered independently in[29], motivated entirely by AI safety, under the almost-equally-unfortunate name "homogeneous ultracontributions". Can we unify these, perhaps as non-empty *topologically-compact* $\perp$-closed convex sets of subprobability measures?
   + Can we incorporate this more general semantics into our answers about compositional stochastic hybrid systems and quantitative bounds?

The group of participants at the workshop also generated several new questions, such as:

1. **Knowledge representation and version control** via "path-dependent" types and a **kind of finite types**: can we construct a type theory in which instances of functional database schemas can be easily represented (by quantifying over finite types, which correspond to sets of row IDs), and which have a semantics in causal preorders, using a form of colimit completion to represent potentially-conflicting states of version-control of such data?

2. **Commutative monad of probability, nondeterminism, and partiality**: regarding the final question above, these monads are not commutative, which is very inconvenient for probabilistic programming. Can we use "variable names" (along the lines of the countably infinite rose tree used to define Ω in[9]) to define a variant which is commutative, and additionally more closely resembles structural causal models?

3. **Functorial boxes in multiple categories**: is there a combinatorial recipe for constructing a multiple-categorical diagram language that bridges between different multiple categories (with certain multiple functors between them) by constructing all the "corner" (and "face") generators?
4. **Double categories for branch-and-bound reasoning**: can we add a multiple-categorical "dimension" to a monoidal category to track relational inclusions (like bicategories of relations but with metric/quantale structure) and use this to define branch-and-bound certificates as double-categorical diagrams?
5. **Hybrid doctrines of systems and specification theories**: can we use fibrational methods to simultaneously define specifications and systems, and develop generic constructions of "hybrid systems" (one type of system fibred over the other), to easily hybridise many different modelling languages?
6. **Outcome logic in partial Markov categories**[12]: what kind of outcome logic[58] or probabilistic verification logic (e.g. [44]) is the best suited? could this be a good semantics for abstract states (in the sense of abstract interpretation[2])?
7. **Global safety from local safety**: can we use a Grothendieck construction to construct global safety proofs in a composite system from safety proofs of the components? is this related to rely/guarantee?

Appendix A.1 Initial list of modelling languages we would like to unify

1. **Differential equations**
   (a) ordinary (ODEs)
   (b) partial (PDEs)
   (c) stochastic (SDEs, SPDEs)
   (d) random (RODEs, RPDEs)
   (e) jump-diffusion
2. **Markov processes**
   (a) discrete-time Markov chains (DTMCs)
   (b) continuous-time Markov chains (CTMCs)
   (c) Markov decision processes (MDPs)
   (d) Markov automata (MA)
   (e) open games
   (f) (open?) mean-field games
3. **Hybrid systems**
   (a) Generalised stochastic hybrid systems (GSHS)
4. **Petri nets** (PNs)
5. **Probabilistic models**
   (a) **probabilistic graphical models** (PGMs)
      i. Bayesian networks (BNs)
      ii. structural causal models (SCMs)
      iii. Markov random fields (MRFs)
   (b) corecursive programs in a functional **probabilistic programming language** (PPL)
      + including, notably, autoregressive large language models (LLMs)
   (c) **probabilistic logic programs** (ProbLog)
   (d) **score-based generative models** (SBGMs)
      + including, notably, diffusion models

Appendix B Game theory analysis — modelling assumptions

The crucial considerations regarding the balance of accident risks against misuse risks and economic opportunity costs in a strategic setting already appear (in terms of Nash equilibrium structure) in the simplest possible game-theoretic framework, a 2-player normal-form symmetric bimatrix of payoff utilities for two strategies ("Saf" and "Main"):

	Player A chooses safe design	Player A chooses mainstream
Player B chooses safe design	A: $U(\text{Saf}, \text{Saf})$, B: $U(\text{Saf}, \text{Saf})$	A: $U(\text{Main}, \text{Saf})$, B: $U(\text{Saf}, \text{Main})$
Player B chooses mainstream	A: $U(\text{Saf}, \text{Main})$, B: $U(\text{Main}, \text{Saf})$	A: $U(\text{Main}, \text{Main})$, B: $U(\text{Main}, \text{Main})$

A simple model of the four expected utility variables is based on the following assumptions:

1. Ultimately, there are eight possible unmixed outcomes, spanned by the 3 binary variables

$$\{\text{LoseRace}, \text{WinRace}\} \times \{\text{SafeDesign}, \text{Mainstream}\} \times \{\text{Accident}, \text{Aligned}\}$$

2. Accident is a Bernoulli random variable whose probability is reduced by SafeDesign:

$$\mathbb{P}(\text{Accident}|\text{SafeDesign}) < \mathbb{P}(\text{Accident}|\text{Mainstream})$$

3. For simplicity, we assume that there are values $0 < \alpha \leq 1$, $0 < \beta \leq 1$, such that

$$U(\text{SafeDesign} \wedge \cdots) = \alpha \cdot U(\text{Mainstream} \wedge \cdots) + (1 - \alpha) \cdot U(\text{Accident})$$

   (i.e. α is the fraction of Mainstream economic value/utility that can still be gained via SafeDesign), and

$$U(\text{LoseRace} \wedge \text{Mainstream} \wedge \text{Aligned}) = \beta \cdot U(\text{WinRace} \wedge \text{Mainstream} \wedge \text{Aligned})$$

   (i.e. β is the fraction of value that is retained even if one loses the race to an unleashed opponent).
4. Because SafeDesign methods should be used to end the acute risk period as soon as possible, the economic loss α (from restricting scaling to only SafeDesign systems) is only during an initial period of T years, when economic returns are a factor of α_0 less, accounted for with an annual discount factor of γ:

$$\alpha = \frac{\int_0^T \alpha_0 \gamma^t \mathrm{d}t + \int_T^\infty \gamma^t \mathrm{d}t}{\int_0^\infty \gamma^t \mathrm{d}t} = \alpha_0(1 - \gamma^T) + \gamma^T$$

   We use the US FEDFUNDS rate at time of writing (5.33%) to set the default discount factor, $\gamma = 1/(1 + 5.33\%)$, to reflect the time-preference of large-scale capital flows.
5. Because a SafeDesign (in our vision/definition) would be multilateralist, we assume

$$U(\text{LoseRace} \wedge \text{SafeDesign}) \geq 99\% \cdot U(\text{WinRace} \wedge \text{SafeDesign})$$

6. If both players have the same strategy, WinRace will be 50%.
7. $\mathbb{P}(\text{WinRace}|\text{Saf}, \text{Main})$ is very low, but if this event takes place, it implies a SafeDesign:

$$\text{WinRace} \wedge (\text{Saf}, \text{Main}) \Rightarrow \text{SafeDesign}$$

8. If A and B both play Saf, then the outcome will be a SafeDesign.

$$(\text{Saf}, \text{Saf}) \Rightarrow \text{SafeDesign}$$

9. In case of Accident, WinRace or LoseRace, SafeDesign or Mainstream... don't matter.
10. Without loss of generality (since utilities are invariant under affine transformation), we let

$$U(\text{Accident}) = 0$$
$$U(\text{WinRace} \wedge \text{Aligned} \wedge \text{Mainstream}) = 1$$

For the purposes of Figure 2 we have selected the following parameters:

$$\begin{aligned}
\mathbb{P}(\text{Accident}|\text{SafeDesign}) &= 0.6\% \\
\mathbb{P}(\text{Accident}|\text{Mainstream}) &= 50\% && (\text{“minimally confident”, per Wasserstein}^{11}) \\
\mathbb{P}(\text{WinRace}|\text{Saf}, \text{Main}) &= 5\% \\
\beta &= 10\% \\
\gamma &= \frac{1}{1+5.33\%}
\end{aligned}$$

The remaining parameters of the model (α_0 and T) are the x and y axes of Figure 2, respectively.

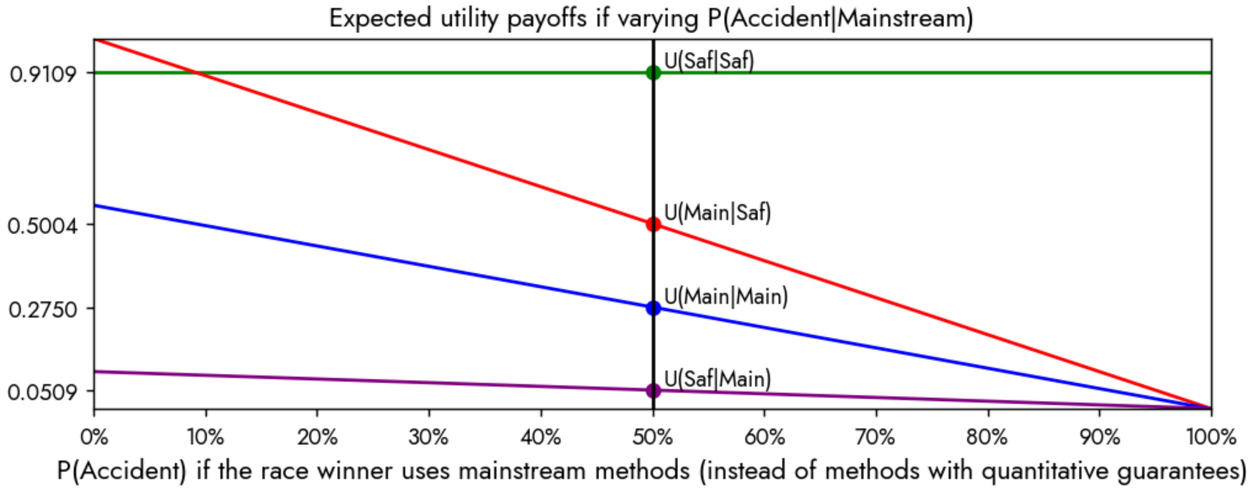
At the specific point marked "Reality after a successful programme", we have $T = 10$ years and $\alpha_0 = 16.9\%$, which implies $\alpha \approx 66\%$. This yields the outcome payoffs:

	WinRace	LoseRace
Aligned ∧ Mainstream	1.00	0.10
Aligned ∧ SafeDesign	0.66	0.65
Accident ∧ Mainstream	0.00	0.00
Accident ∧ SafeDesign	0.00	0.00

and the expected utility variable values:

		Other	
		Saf	Main
Self	Saf	0.66	0.08
	Main	0.51	0.28

Figure 5: The sensitivity of the expected utility variables at ($\alpha_0 = 16.6\%$, $T = 10$ years) to $\mathbb{P}(\text{Accident}|\text{Mainstream})$, at a "distant future" discount rate of $\gamma = 1/(1+1\%)$[52]. The structure of the bimatrix game changes where the payoff lines cross, from Prisoner's Dilemma at the left to Stag Hunt in the middle to No Conflict at the extreme right.



[11] It is a common mistake to implicitly assume a privileged reference measure μ on a set like {Accident, Aligned}, which is also necessary to argue from "maximum entropy" or "minimum information" that one should privilege a probability measure ν, like so:

$$\nu = \arg\min_\nu D_{\mathrm{KL}}(\nu \mid \mu) = \arg\min_\nu \int_\Omega \mathrm{d}\nu(\omega) \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\omega)\right)$$

Counting measure can be justified by permutation-invariance, but there is no reason to assume that a set like this is permutation-invariant. *However*, if we have a pseudometric on the outcome space Ω, we can instead argue for the Wasserstein barycenter as the "least confident" measure,

$$\nu = \arg\min_\nu \arg\max_\mu W_1(\nu, \mu)$$

Given a utility function $U : \Omega \to \mathbb{R}$, we do indeed have a pseudometric on $\Omega = \{\text{Accident}, \text{Aligned}\}$, namely $d(x, y) = |U(x) - U(y)|$, and it can be shown that **the "least confident" Wasserstein barycenter puts 50% probability mass on each** of {Accident, Aligned}.