Programme Discussion Report

David ‘ davidad’ Dalrymple | Compositional World Modelling | TA1.1 | March 2024

Programme Discussion Report

One of the key underlying principles of ARIA’s solicitation is fair, open and transparent competition, as such we are publishing a summary of the outputs of programme discussions.

SECTION 1: Programme discussion overview

Programme / Programme Director	Safeguarded AI / David ‘davidad’ Dalrymple
Date	04/03/2024 - 05/03/2024
Location	Deaf Cultural Centre 100 Ladywood Rd, Birmingham B16 8SZ
Goal	Continue to progress in brainstorming and scoping concrete research questions for TA1.1
Pre reads	Pre-read — Some Concrete Research Questions For The Programme Discussion [PDF]

Background

Much of the value of the Safeguarded AI programme derives from constructing a scalable and highly expressive language for real-world dynamics, goals, safety properties, and neural controllers. If successful, such a language will serve as the mathematical workbench for collaborative teams of human stakeholders, domain experts, and AI assistants to formalise scientific models, engineering specifications, and AI solutions pertinent to real-world problems. All the other technical areas in the programme depend in some way on the definition of syntax and semantics for this language: TA1.2 would be implementing it, TA1.3 implementing a user interface for it, TA2 training AIs to perform tasks within it, and TA3 using it to solve specific problems.

Current knowledge-representation schemes for digital twins (e.g. Modelica, DTDL) are not adequate to express or version-control probabilistic models of real-world systems, let alone proofs that neural controllers for those systems satisfy certain probabilistic-logical specifications. The programme thesis evinces a conviction that, in light of the rapidly accelerating code-generation and proof-automation capabilities of frontier AI models, there are latent possibilities within category-theoretic algebras of system composition to define radically more practically-valuable modelling languages. This could be analogous to the revolutionary role played by the set-theoretic algebra of relations in the formulation of modern database theory in 1970.

The purpose of this discussion was to continue progress from the programme workshop in December 2023 regarding brainstorming specific research directions that seem tractable and for which solutions might plausibly be critical for a next-generation language of probabilistic models, specifications, and/or proofs.

The goals of the invitee selection were:

- to curate the nucleation cluster for a new R&D community;
- centred in applied category theory, but with bridges to formal verification, AI safety, and machine learning research;
- centred in the UK, but with bridges to other leading groups worldwide;
- optimising for geographic diversity, gender diversity, and institutional diversity as much as feasible without compromising topical focus or quality.

The agenda was designed to foster the rapid exchange of ideas and to facilitate efficient progress in generating actual new mathematical content. The format was slightly evolved but closely based upon the success of the programme workshop in December 2023. The core of the format is multiple multi-hour breakout sessions in which participants self-select into groups around topics which any participant can nominate. Before each breakout session, the potential topics are narrowed down to 2-4 through real-time, closed-eyes approval voting. Then each group goes into a dedicated room with a whiteboard and a dedicated facilitator who keeps time and occasionally interjects with metacognitive suggestions to keep participants engaged and to keep the conversation on track. The hoped-for output at the end of *each* breakout group is a consensus on the tractability of one research problem that could be undertaken by one or a few students over the course of 3—24 months. In addition to these breakout sessions, the agenda consisted of several lightning talks, brief context-setting at opening and closing, and a longer talk from Yoshua Bengio with a Q&A session.

Key discussion points

The following topics were discussed during breakout sessions:

1. **Commutative monads for imprecise probability**. We have good commutative monads for nondeterminism and probability separately, and excitingly, in the past few months the programme community has already identified a commutative monad that combines them. It has recently been shown that a non-normalized monad of measures is useful in posterior inference; we discussed whether this can be combined with imprecise probability. One tractable direction is to use the Reader monad transformer applied to the existing non-normalized monad of measures, to supply additional, named, nondeterministically uncertain variables. Questions remain about how to implement names, what quotients may be appropriate, and whether this really is beneficial in practical implementation.
2. **The interfaces with machine learning (TA2)**. Given what we know so far, it was elaborated that a world-model and spec together interface with an ML system in three ways: first, potentially, as a kind of input "prompt"; second, via a formal verifier—which should ideally be differentiable, so that a gradient signal of "verifiability" can be used as a loss function to steer ML optimization; and third, being *refined* by an ML system. This underscored the need for expressive, compositional specs and proofs.
3. **Double-fibrational approaches** to hybrid systems and specs, and potential connections to rely/guarantee reasoning, outcome logic, and temporal type theory. In this approach, one axis of the double category is interfaces or boundaries and their morphisms (like abstractions or functions of the boundary-layer state spaces), and the other axis has systems for morphisms.
4. **Presenting monoidal double categories of modal logics** that include probabilistic temporal specifications. It was again noted that these would likely take semantics in the double category of spans or something similar. A suggested stretch goal, estimated to take an additional 3 person-years, is to define how to extract a notion of time from a notion of system that doesn't have a pre-existing time axis.
5. **A "big tent" semantics in a double category V-Rel of enriched relations**. The most critical feature of systems to model in this programme is imprecise reachability

probabilities. If it weren't for the probabilistic aspect, imprecise reachabilities would be ordinary relations, which are organised in the double category Rel. So, can we find a quantale V such that V-Rel is a good semantic universe that incorporates the probabilistic aspect?

6. **Compositional specs beyond reachability**. If systems have semantics as spans of trajectories (behaviours), and specs have semantics as relations of trajectories, then the spec semantics can be extended to spans, and a system's satisfaction of a spec can be defined as a (2-)morphism of spans. It was noted that a pathway to "tractable compositional verification" might begin with defining such a specification language as a theory in a doctrine of dynamical systems such that known verification techniques are subsumed (such as modal mu-calculus, refinement types, rely/guarantee, etc.).

The following topics were also raised, but were not the formal subject of a breakout discussion:

1. **Reconfigurable port-Hamiltonian systems**, in which hybrid switching transitions could modify the port graph
2. **Version-controlled compositional knowledge representation**, using sigma-types indexed by FinSet
3. **Extending stochastic dynamic coloured Petri nets to infinite-dimensional state spaces**
4. **The "sheafy" perspective on dynamical systems**, as exemplified by Temporal Type Theory, and whether this can be extended to encompass stochasticity

Outcomes

The programme discussion met its goals of continuing to identify tractable research directions in TA1.1 of the programme. Perhaps most useful is the emerging consensus that systems and specifications should be organised into two different (double) categories which each take semantics in something like spans (e.g. spans weighted by a quantale), with proof certificates taking semantics ultimately in the morphisms of such spans.

Having confirmed the ripeness of these directions, the next step is a call for proposals in TA1.1.

SECTION 2: Agenda

Day 1 Agenda | [04.03.2024]

Time	Agenda Item
10:00 - 11:00	Introduction to programme discussion and programme presentation
11:15 - 13:30	Breakout session 1
14:30 - 15:45	Lightning talks: *Owen Lynch - Formal and Informal Collaboration* *Mario Román - Bayesian networks form the free Markov multicategory* *Sam Staton - Compositional Imprecise Probability* *Elena Di Lavore - Effectful Trace Semantics* *Mirco Giacobbe - Progress on proof certificates* *Wen Kokke - Vehicle: Bridging the Embedding Gap in the Verification of Neuro-Symbolic Programs*
16:00 - 16:45	ARIA discussion & Q&A
16:45 - 17:00	Day 1 closing remarks

Day 2 Agenda | [05.03.2024]

Time	Agenda Item
09:30 - 09:40	Welcome & briefing
09:40 - 11:20	Breakout session 2
11:30 - 12:20	Yoshua Bengio - prepared remarks and Q&A
13:30 - 17:15	Breakout session 3
17:15 - 17:35	Feedback
17:35 - 17:45	Day 2 closing remarks