I. The MathZero Project

II. AI Safety, AI Architectures, and AI Consciousness

David McAllester

Friday, November 10, 2023

The MathZero Project

This has the ultimate goal of accomplishing for “the game of mathematics” what AlphaZero did for Chess and Go.

A much more modest goal is improving the level of automation in formal verifiers for general mathematics.

The most popular verifier among mathematicians is the LEAN system. The math library for LEAN (mathlib) currently contains 70,000 formal definitions and 127,000 formal proofs.

I think it is safe to say that attempts to use deep learning to improve automation in LEAN have not yet lead to significant productivity improvements.

An Approach to MathZero

Step I. Get the foundations right. Specify an appropriate variant of **dependent type theory**. The version considered here is different from that underlying LEAN.

Step II. Construct an effective verification system for "declarative proofs" rather than "syntactic proofs". Eliminate the concept of "tactic".

Step III. Continuously improve the verifier until you only need to state a desired theorem. Self-play is replaced by a curriculum of increasingly difficult problems.

Getting the Foundations Right

The set of theorems of mathematics is defined by the nine axioms of Zermello Fraenkel set theory with the axiom of choice (ZFC).

However, the formulas of ZFC are very different from the natural language (e.g. English) used by mathematicians.

It is the difference between assembly code (ZFC) and a strongly typed high level language (dependent type theory).

As with programming languages, variants of dependent type theory can differ.

Getting the Foundations Right

Unlike LEAN, here the type theory is specified by defining **the meaning** of type expressions.

Type expressions denote sets or classes.

The type `set` denotes the class of all sets.

$\sigma \times \tau$ denotes the set (or class) of pairs (x, y) with $x{:}\sigma$ and $y{:}\tau$.

The **dependent type** $(x : \sigma) \times \tau[x]$ denotes the set (or class) of pairs (x, y) with $x{:}\sigma$ and $y{:}\tau[x]$.

For example $(s : \texttt{set}) \times ((s \times s) \rightarrow s)$ denotes the class of all magmas.

Types and Abstraction

There is no natural or canonical ordering on an abstract set.

There is no natural or canonical point on a geometric circle.

There is no natural or canonical basis (coordinate system) for a vector space and, relatedly, no natural or canonical inner product operation on the vectors of a vector space (an inner product is the same data as an isomorphism between a vector space and its dual).

The Problem with Category Theory

Category theory does not recognize the role of type theory in defining “functor”, “canonical object”, or “same data”.

`Topology` → `Group` is the class of all “functors” (**well-typed** function expressions) from topological spaces to groups.

In general for $x : \sigma$ there is no canonical element of $\tau[x]$ if there is no **well typed** function $F : (x : \sigma) \rightarrow \tau[x]$.

The objects in σ and τ contain the same data if there are **well typed** functions $F : \sigma \rightarrow \tau$ and $G : \tau \rightarrow \sigma$ with $G(F(x)) = x$ and $F(G(y)) = y$.

Mathematical Objects have Symmetry Groups

Consider the bag-of-words abstraction of a document (a mapping from words to the number of times they occur).

Clearly, the bag of words has lost the information of the order of the words.

In dependent type theory we have that, for an abstract alphabet $\mathcal{A}$, there is no well-typed function

$$G : \mathrm{BagOf}(\mathcal{A}) \to \mathrm{SequenceOf}(\mathcal{A}).$$

More generally, in dependent type theory **objects have symmetries** and **well-typed functions cannot break symmetries**.

Types and Isomorphism

Dependent type theory associates every type with a notion of isomorphism and supports the substitution of isomorphics.

For $x, y : \sigma$ I will write $x =_\sigma y$ to mean that x and y are isomorphic as σ.

$$\frac{\begin{array}{l}\Gamma \vdash F{:}\sigma \to \tau \\ \Gamma \vdash u =_\sigma v\end{array}}{\Gamma \vdash F(u) =_\tau F(v)}$$

Getting the Foundations Right

The type

$$(x{:}\,\sigma)\ \texttt{such that}\ \Phi[x]$$

denotes the set (or class) of $x:\sigma$ satisfying $\Phi[x]$.

Unlike LEAN, here the same object can have many types — an Abelian group is a group.

Unlike LEAN, but as in Tarskian semantics, equality means equality — here equality is not defined as an inductive type.

Declarative Proofs

Declarative proofs use only a small number of constructs.

Perhaps only `Proof`(Φ,proof), `LetBe`($x{:}\sigma$, proof), `Suppose`(Φ,proof) and `NoteThat`(Φ).

The system must verify the proof leaves of the form `NoteThat`(Φ).

Eliminating tactics greatly simplifies the search for proofs.

Human proofs do not call tactics.

There is a continuum of strength for declarative verifiers with no upper limit.

Supporting Declarative Proofs

Before training deep models I am adapting techniques from SMT solvers such as Z3.

SMT solvers use highly effective inference “algorithms” such as unit propagation and congruence closure.

SMT methods need to be adapted to dependent type theory.

For example congruence closure for the isomorphism equivalence relation.

Release Date Target

I am targeting September next year (2024) for the release of a competitor to the LEAN verification system.

Part II

AI Safety

AI Architecture

AI Consciousness

Safety: The Alignment Problem

The Alignment problem is that of giving an artificial general intelligence (AGI) a mission or purpose in alignment with human values.

This can be phrased as finding a solution to the principal-agent problem for AGI agents.

White-Hats, Red-Hats

A white-hat team designs a safety system.

A red-hat team looks for vulnerabilities.

We need both.

White Hat: The Advobot Restriction

A personal advobot is an AI advocate for a particular person X where the advobot's fundamental goal is given as "within the law, pursue fulfilling the expsed requests of X".

The advobot restriction is that AGI systems be legally limited to personal advobots.

The term "AGI" needs to be incorporated into law and given an evolving interpretation by the judicial system.

White Hat: Safety Features of the Restriction

- The advobot must act within the law. Society can limit all advobots by changing the law.
- The advobot mission transfers moral responsibility from the advobot to its master.
- There is a large society of advobots — one per person — each with a different mission. This limits individual power.
- The advobot mission seems clearer that other directives such as Asimov's laws or Yudkowsky's coherent extrapolated volition.
- The advobot restriction preserves human free will.

Red Hat: Consider Large Language Models (LLMs)

Much of the literature on AI safety assumes that we can give an AI a goal such as “make as many paperclips as possible”.

But large language models (LLMs) are not even “agentive” (explained below).

LLMs are trained to mimic people. People do not have clear objectives and do not always do what they are told.

Large language models are subject to the “Waluigi effect” where they flip to pursuing the very opposite of what they are told.

Agentive AGI

An AGI system is “agentive” if it takes actions in puruit of a goal.

Many systems can be decribed as taking actions in persuit of a goal. But an AGI is agentive if its potential actions include all the kinds of actions that people can take. For example legal filings of all kinds.

Current LLMs are not agentive.

The Waluigi Effect

Waluigi is the evil twin of Luigi in Mario Brothers.

The Waluigi effect occurs when an LLM holds two interpretations of its own statements — one genuinely cooperative and one deceptively cooperative.

When modeling humans both interpretations exist.

If the LLM reveals deception, the deception interpretation sticks.

Every turn of the dialogue has a chance of revealing deception.

White Hat: Constitutional AI

Constitutional AI is an attempt to provide a mission statement (or “constitution”) to LLMs.

Constitutional AI has been show to work to some extent but is not included in GPT4 which instead uses reinforcement learning with human feedback (RLHF).

Ultimately it seems clear that we need to be able to specify missions.

Constitutional AI: Harmlessness from AI Feedback

Bai et al ArXiv 2212.08073 [Anthropic]

Memory Architectures

For both safety and performance reasons I believe strong AGI systems will be based on read-write memory architectures.

In a memory architecture a “CPU” works with an external memory in a manner analogous to a von Neumann machine.

We might have a **transformer CPU** where the transformer context is analogous to registers in a classical CPU.

Items can be loaded from memory into the CPU context and written from the CPU context into memory.

Related Literature

There are hudreds of papers on enlarging the transformer context.

There are hundreds of papers on retrieval of documents or entities in a knowledge graph.

There are papers that maintain indefinite term state in toy sandbox domains.

I have not seen papers proposing read/write memory that includes internal thoughts over general language. Send me pointers if know of some.

Performance Advantages of Memory Architectures

The memory acts as an essentially infinite context with memory retrieval playing the role of the attention mechanism of a transformer but over all of memory.

The memory can be directly extended. The machine can read and remember today's newspaper.

The machine can use internal chain-of-thought processing involving reads and writes to memory.

Safety Advantages of Memory Architectures

We want to know what an agent believes.

We want to know the agents goals.

We want both of these things to be visible in the memory.

Interpretability (Opening the Black Box)

We should be able to engineer the memory such that memory entries are either literally textual statements, or have a rendering as text, and where the textual representation is faithful to meaning assigned by the machine.[1]

By observing the bandwidth to memory we can observe the "thought process" of the machine.

We can also edit the memory to maintain the quality of its information, or control the beliefs of the machine.

[1] For example, the machine's notion of entailment between memories is in correspondence with human entailment judgements between their textual representations.

Mission Statements (Fundamental Goals)

Fundamental goals are axioms. They do not follow from, and are independent of, world knowledge.

An axiomatic **and immutable** mission should be built into the CPU.

The Advobot Restriction

A personal advobot is an advocate for a particular person X whose fundamental goal is given as “within the law, pursue fulfilling the expressed requests of X”.

The advobot restriction is that AGI be limited to personal advobots.

Defining AGI

Legally limiting AGI to advobots requires some legal interpretation of “AGI”.

AGI is of course hard to define.

However, many legal terms are hard to define. Consider “intent”, “bodily harm”, or “assault”.

Perhaps we can simply use the term “AGI” in legal discourse and leave its interpretation open to an evolving legal process.

A Possible Legal Definition of AGI

An Artificial general intelligence (AGI) is a a non-biological computational system possessing the following abilities at least at the level of a normal person.

- The ability to converse in language on topics familiar to most people.

- The ability to understand counterfactuals, the consequences of actions, and to track the state of the world as events unfold.

- The ability to understand the mental states of others as a part of the evolving state of the world.

- The ability to pursue goals through actions. This includes, but is not restricted to, telling people things, asking people to do things, engaging in financial transactions, and filing legal documents.

- It has unspoken thoughts that are remembered and that can be used to improve understanding or to plan actions.

- It remembers what it is told, what it reads, what it says, and its own thoughts.

A Possible Legal Definition of Consciousness

Any system that passing the legal definition of an AGI is to be considered legally conscious.

Memory is a particularly important criterion for consciousness.

Defining Truth

While it may be possible to edit the beliefs of an advobot, one might want legal protection for truth in advobot beliefs.

This would involve the ability to legally interpret “truth”.

But the legal system has always had to judge truth.

END