Abstract Algebra for AI Safety?

Dovetail Research is offering a paid research fellowship for UK-based researchers interested in uncovering the mathematics behind agents. Here's one problem you could work on.

Let $\mathcal{A}$ be an algebraic structure that could reasonably be said to describe the world, such as a dynamical system, a Lie group, or an automaton. AI systems necessarily learn coarse-grained representations of the world. Thus, one way we could think of their internal world models is as **quotient objects** of the $\mathcal{A}$-algebra—that is $\mathcal{A}$-congruence relations, or images of $\mathcal{A}$-homomorphisms. For a given $\mathcal{A}$-algebra the collection of all such quotient objects form a lattice. We wish to investigate whether the structure of such lattices can tell us useful information about the world models that AIs are learning.

Directions of inquiry

- Which type of algebra do ML systems learn?
- Is the quotient object assumption accurate?
- Which algebras have had their lattices studied?
- What kinds of lattice structures would tell us something about AI safety?
- How can we discern how far up the lattice an AI is?

Motivating example

Will all sufficiently advanced AIs learn Newtonian mechanics? If the lattice has a "bottleneck"—a maximal anti-chain of width 1—then any AI system should be expected to converge on that world model as their model granularity passes through that level.

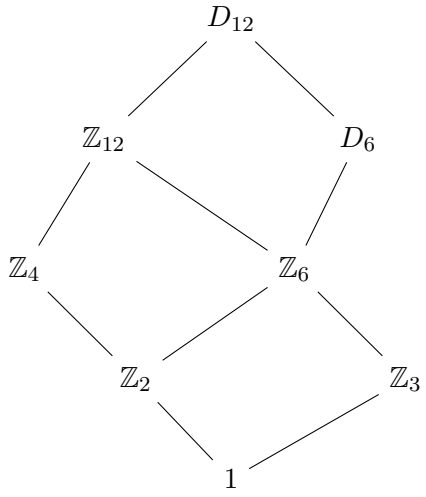

Fig. 1 The lattice of quotient groups (up to isomorphism) of the dihedral group D_{12}. A quotient group acts like a coarse-grained "theory" or "model" of the parent group.

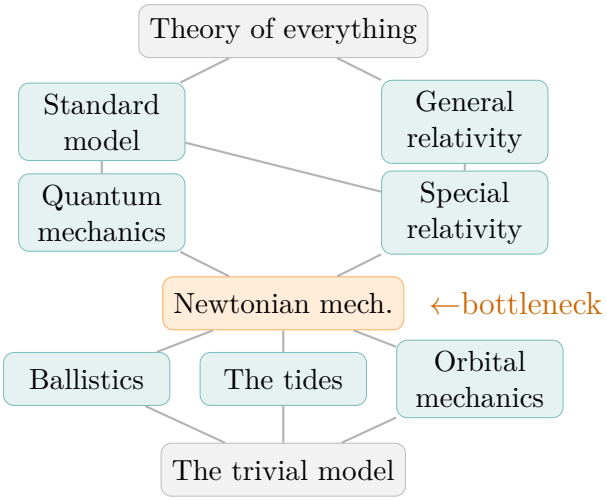

Fig. 2 A lattice of different physical theories across a range of coarse-grainings. How accurate is this analogy?

What does this have to do with AI Safety?

ML systems are notorious for having undesirable behaviour "out of distribution". This is in large part due to our lack of theoretical understanding of what they learn. If we can find any kind of correspondence between the structure of these lattices and the internals of ML models, that may give us the ability to predict and control this behaviour.

If you have experience in maths, physics, or computer science and are interested in working on problems like this, apply using the QR code. Rolling acceptance until application closes on May 17.

This job is part of an Advanced Research + Invention Agency-funded project.

To see more problems or apply scan here or visit `dovetailresearch.org`