Theoretical Computer Science 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.

Touchette and Lloyd (2000) proved a clean result linking **optimization** and **world-modelling**. Consider an environmental state (random variable) X, transformed into a next state Y under the influence of an action A and dynamics characterised by a conditional distribution $P(Y|X,A)$. We can analyse the conditional distribution over actions $P(A|X)$ (the 'policy'). Let H denote Shannon entropy and I denote mutual information. Define the entropy reduction as $\Delta H := H(X) - H(Y)$. Let ΔH^{Max}_{Blind} be the maximum entropy reduction achievable by any *blind* policy (where A is independent of X so $P(A|X) = P(A)$), maximized over input distributions and blind action distributions: $\Delta H^{Max}_{Blind} := \max_{P(X),\,P(A)} \Delta H$. The theorem states

$$\Delta H \leq \Delta H^{Max}_{Blind} + I(X;A).$$

Intuitively: each extra bit of entropy reduction beyond what the dynamics permit "for free" requires at least one bit of information about the environment.

We wish to develop an analogous inequality where the resource is not Shannon entropy, but **Kolmogorov complexity** (algorithmic information). Let $K(\cdot)$ denote prefix-free Kolmogorov complexity of a string. Let us define x, y, a as strings and let the dynamics be given by a deterministic function $y = f(x,a)$. We are interested in whether there exists a similar relationship between the change in Kolmogorov complexity $\Delta K := K(x) - K(y)$ and the *algorithmic mutual information* between x and a, which is defined as:

$$I_K(x;a) := K(x) - K(x|a^*)$$

where a^* is the shortest program which prints a.

Directions of inquiry

- What is the right notion of a "blind" baseline when the quantity of interest is Kolmogorov complexity rather than entropy?
- Kolomogorov complexity inequalities are often true only up to additive constants. How should we handle these so that the theorem is robust?
- Can we relate the resulting bound to compression or generalization phenomena of learned world models in AI systems?

Why complexity, not entropy?

Kolmogorov complexity captures *description length* of individual outcomes, not just average-case uncertainty. This may better match settings where agents exploit rare but highly structured regularities, or where the "world model" is best viewed as a program rather than a probability distribution.

What does this have to do with AI Safety?

If we can link how much *algorithmic* structure an agent can impose on its environment with its algorithmic information content about that environment, we get a new lever for arguing that certain kinds of capability require internal representations that are informative about the world. This might help us understand when certain capabilities imply the ability for dangerous and agentic 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