New TFI Research on Structural Risks: The Reasoning Under Uncertainty Trap

We are pleased to announce the release of a new TFI research article, written by Senior Research Scientist Dr. Toby D. Pilditch.

The long-form report details a novel, current AI misuse risk, which is then compounded by surrounding structural risks. Critically, the report outlines how the complex structure of these interlinked structural risks can not only make detection difficult, but have feedback properties that turn an otherwise additive risk impact into a non-linear growth of potential harm.

The short abstract below provides an overview:

"This report examines a novel risk associated with current (and projected) AI tools. Making effective decisions about future actions requires us to reason under uncertainty (RUU), and doing so is essential to many critical real world problems. Overfaced by this challenge, there is growing demand for AI tools like LLMs to assist decision-makers. Having evidenced this demand and the incentives behind it, we expose a growing risk: we 1) do not currently sufficiently understand LLM capabilities in this regard, and 2) have no guarantees of performance given fundamental computational explosiveness and deep uncertainty constraints on accuracy.

This report provides an exposition of what makes RUU so challenging for both humans and machines, and relates these difficulties to prospective AI timelines and capabilities. Having established this current potential misuse risk, we go on to expose how this seemingly additive risk (more misuse additively contributed to potential harm) in fact has multiplicative properties. Specifically, we detail how this misuse risk connects to a wider network of underlying structural risks (e.g., shifting incentives, limited transparency, and feedback loops) to produce non-linear harms.

We go on to provide a solutions roadmap that targets multiple leverage points in the structure of the problem. This includes recommendations for all involved actors (prospective users, developers, and policy-makers) and enfolds insights from areas including Decision-making Under Deep Uncertainty and complex systems theory. We argue this report serves not only to raise awareness (and subsequently mitigate/correct) of a current, novel AI risk, but also awareness of the underlying **class** of structural risks by illustrating how their interconnected nature poses twin-dangers of camouflaging their presence, whilst amplifying their potential effects."

Here at TFI we believe structural risks to not only be an essential, yet mostly overlooked area of AI safety research. This report serves to illustrate by example not only how this class of risks can pose current potential for outsized societal-scale harms, but also demonstrate possible in-roads for evaluations, mitigation, and policy solutions.

Read the full research article here: The Reasoning Under Uncertainty Trap.