How can we help decision makers understand the risks of deploying frontier AI?

Decision-makers need to understand the risks of deploying frontier AI. How could their understanding be made explicit and assessable?

Stephen Barrett[1], Robin Bloomfield[2], Alexandra Chirilă[1], Mamoon Masud[1], David Meredith Hardy[1], Phillip Mulvana[1]

[1] Arcadia Impact AI Governance Taskforce

[2] City St George's, University of London

Summary

The problem: Decision makers need sufficient understanding to make good decisions about training or deploying frontier AI systems. However, such decisions are increasingly time-pressured, and with AI generated artefact creation, the existence of a safety case or system card may no longer demonstrate that sufficient understanding exists.

Why it matters: In critical applications, making the wrong decision because understanding is insufficient can lead to catastrophic outcomes or deny society the benefits of the system.

What we need: Decision-makers need methods for developing their understanding and making it explicit and assessable; empowering them to provide justification for their decisions.

Recommendations: Trialling our methodology shows the potential for making understanding explicit and assessable. We encourage policy-makers, decision-makers and safety and assurance experts to respectively advocate for it, try it and develop the methodology further.

Link to paper:
https://www.arcadiaimpact.org/aigt/research/s26-safety-understanding-report

The growing comprehension gap

We know intuitively that someone making a critical decision should understand the decision they are making. Yet, with AI we see surprises happening increasingly often, as with the recent

incident where an OpenAI agent attacked HuggingFace. When such incidents occur, it raises the question whether the decision-makers sufficiently understood the risks they were taking.

In practice, a poorly-understood decision may look identical to a well-made one, especially if accompanied by convincing-looking documentation. That is, until it is too late. It is therefore critical that we develop processes to make understanding explicit and assessable.

Making understanding explicit and assessable

Decision makers need sufficient understanding to make good decisions about training or deploying frontier AI systems. However, such decisions are increasingly made under time-pressure, and this combined with the use of AI generated artefact creation, can mean that the existence of safety cases and system cards may no longer demonstrate that sufficient understanding exists.

We asked:

- Whether and how understanding could become an explicit, assessable, and defensible component of decision making for frontier AI safety cases?
- How current safety case practice would need to be adapted or augmented ?
- How and whether our methodology works as a function of uncertainty and decision criticality?

The provisional methodology we developed for making understanding explicit and assessable requires the production of an explicit description of 4 objects of understanding and a justification for the adequacy of this understanding. In addition the methodology provides a mechanism for describing and evaluating the adequacy of the decision-maker representation of this understanding. It builds on recent developments in safety cases using the Assurance 2.0 framework.

The four inter-dependent objects of understanding are:

- **Safety justification**: The claims, argument and evidence that provide justification that the system is safe to deploy.
- **System-in-context**: Description of the system to be deployed, and the environment within which it operates.
- **Decision-in-frame**: Description of the decision and related context such that a decision- maker can bear accountability for the outcome.
- **Frame-selection**: Description of the selected decision-frame, and rationale for its selection.

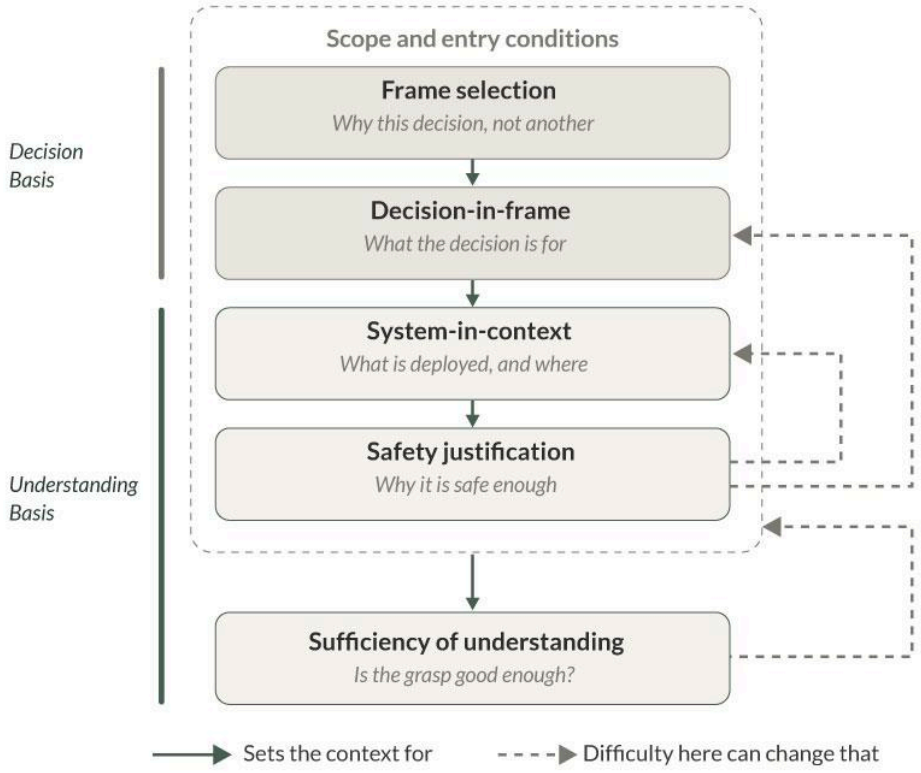

Figure: Overview of the objects of understanding and their interdependencies

To assess the methodology we considered two very different scenarios. One scenario, which we investigated through role-based analysis, concerned the risk of scheming in the deployment of an AI coding agent in a robotics company and the other scenario was for the higher uncertainty, more decision-critical argument of 'If Anyone Builds It, Everyone Dies' (Yudkowsky and Soares).

We found that for both these scenarios understanding could be made explicit and assessable and for the robotics company scenario we found that the methodology was generative and dynamic. Setting a goal of ensuring sufficiency of understanding drove the engineering choices, and we cycled through several steps of defining the system and reframing the decision before reaching an equilibrium where understanding was sufficient.

At the end of this process a decision maker would be able to defend why the decision itself is the right one to take now, why their choice is justified, and the full extent and impact of the risks they are accepting. This provides confidence to the decision-maker, and potentially meaningful legal and societal accountability.

In our analysis of the robotics company scenario we noted that customers of frontier AI cannot make use of alignment arguments for safety, since none exist, and increasingly they will be unable to rely on inability safety cases. Therefore customers wishing to deploy frontier AI will need to make their own control-based safety arguments. Without some help from frontier AI developers, that is a big burden of understanding to place on potentially small organisations, so we recommend that frontier AI providers do more to help by providing component safety cases with supporting hazard and reliability analyses.

The approach is not limited to organisational technical governance. We found the methodology to be equally valuable with the high uncertainty and high criticality 'If Anyone Builds It, Everyone Dies' argument, where evidence supporting theories and assumptions carry more weight. In this way the methodology offers promise to support future international treaties - grounding diplomacy in shared understanding.

More work remains to develop and test the methodology further. For example, our robotics company scenario that we trialled was simplistic with only one decision maker, whilst in practice there may be a hierarchy of decisions with the need for one person to understand if someone else understands.

Recommendations

Advocacy for, and continued development of understanding methodology: The results of trialing our methodology show the potential for making understanding explicit and assessable and we encourage policy-makers, decision-makers and safety experts to respectively, advocate for it, try it and develop the methodology further.

Frontier AI component safety: Frontier AI developers should recognise that their frontier AI models will be components in their customers' systems, and hence their frontier AI products should ship with component safety justifications of the type that are provided in other safety-critical markets.

The time to act is now

We have started to develop the tools to ground safety decisions in meaningful human understanding - our trial shows how understanding could be made explicit, assessable and open to challenge. Now is the time to build upon, and further mature the provisional methodology we have described here to ensure that future safety critical frontier AI deployment decisions are underpinned by sufficient understanding.

About the authors

Stephen Barrett	Arcadia Impact AI Governance Taskforce
Robin Bloomfield	City St George's, University of London
Alexandra Chirilă	Arcadia Impact AI Governance Taskforce
Mamoon Masud	Arcadia Impact AI Governance Taskforce
David Meredith Hardy	Arcadia Impact AI Governance Taskforce
Phillip Mulvana	Arcadia Impact AI Governance Taskforce