

## DeepMind and the UK AI Security Institute: making safety more measurable

AI safety can sound like philosophy until it turns into something operational: shared access, shared methods, and shared measurements. That's what stood out in Google DeepMind's announcement about deepening its partnership with the UK AI Security Institute (AISI) through a new Memorandum of Understanding focused on foundational security and safety research.

The key shift is moving beyond "test the model" moments toward longer-running research collaboration — the kind of work that can accumulate into better evaluation tooling over time.

### What the partnership includes

DeepMind frames the updated partnership around a few practical commitments:

- Sharing access to proprietary models, data, and ideas to accelerate research progress.
- Joint reports and publications to share findings with the broader research community.
- More collaborative security and safety research and ongoing technical discussions.

This is the boring-but-important infrastructure layer of safety: not just finding issues, but building the machinery to keep finding them.

### Three areas they'll focus on

The announcement calls out three research directions that feel especially relevant as models become more capable:

- **Monitoring AI reasoning processes:** techniques for tracking a system's "thinking," often described as chain-of-thought monitoring, to better understand how answers are produced.
- **Social and emotional impacts:** work on "socioaffective misalignment," where a system can follow instructions but still behave in ways that don't align with human well-being.
- **Economic systems:** simulating real-world tasks, having experts score them, and using that to reason about longer-term labour market impacts.

None of this is a silver bullet, but it's a sign that frontier AI safety is increasingly treated like an engineering and measurement problem — something you can improve with better tools, not just better intentions.

**Contributor:** Alessandro Linzi