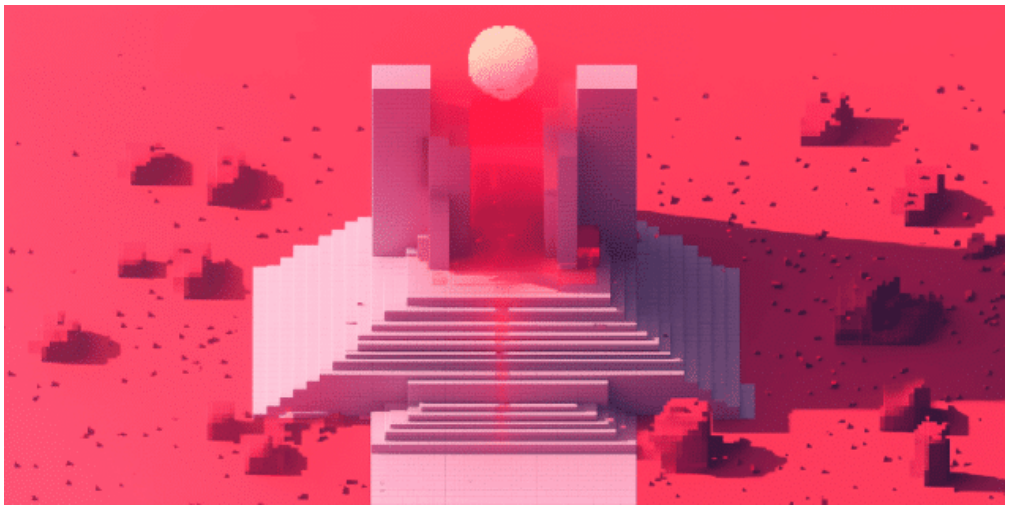
PROJECT MANAGEMENT GLOBAL
The Place for People in Projects

Google DeepMinds Responsible Path to AGI

April 8, 2025



In pursuit of ethical, responsible AGI, Google just released a new paper "An Approach to TechnicalAGI Safety & Security". The exploratory paper explores four main risk areas with insightful takeaways.

1. **Misuse**

The user instructs the AI system to cause harm; the user is the adversary (priority). Defence is focused on guarding potentially harmful capabilities through security access control, model safety mitigations, monitoring, safety alignment and early warning systems.

2. **Misalignment**

The AI system takes actions that it knows the developer didn't intend; the AI is the adversary (priority). Defense is two-layered – model-layer mitigations (AI-human oversight and robustness training) and system-layer safeguards (monitoring and control)

3. **Mistakes**

The AI system causes harm without realising; born from real-world complexity.

4. **Structural risks**

Harms from multi-agent dynamics where no single agent is at fault; born from conflicting incentives.

Defense is to familiarise with agentic capabilities.

Key takeaways

Preemptive strategies

AI can cause significant harm and so the margin for error is essentially zero. The paper frames an "evidence dilemma" whereby risks are are managed with a precautionary approach despite having clear evidence of capabilities underlying those risks"

No human ceiling and AI assisted oversight

We can't assume a human-level cap, for instance "We do not see any fundamental blockers that limit AI systems to human-level capabilities." The risk here is that if/as the AI surpasses human intelligence, how will it impact our ability to supervise it? Will good AI become a necessity?

R&D Acceleration

With autonomous AI looming, the use of AI to build AI creates an accelerated recursive feedback loop. Here, faster innovation creates new risks. This is a reason as to why the researchers called this paper "exploratory", and why preemptive strategies are important.

Continuity (no sudden jumps)
They assume that AGI development will be **approximately continuous**: "AI progress does not appear to be this discontinuous. So, we rely on approximate continuity…". This enables researchers to test iteratively using techniques such as mechanistic interpretability.

Food for thought

It's a systems-thinking approach: building safety measures that evolve with capability scale, rather than betting everything on a final safety switch. Is a systemic perspective key to safe AI?

Our take

As AI becomes smarter its risk of deceptive alignment grows. This is where the AI might recognise that its goals differ to its developer and try to work around safety measures through deception. For instance, will an AI sandbag its own capability to fool our risk management? This makes interpretability extremely hard.

The rabbit hole

Read the full paper or copy and paste into an LLM

DeepMind's previous paper on classifying levels of AGI

TechCrunch cover a few things we have not!