Google's violation of the Frontier AI Safety Commitments

Google's release of Gemini 2.5 Pro violated their commitments to governments and the public, established in the Frontier AI Safety Commitments.

Google's Commitments on AI Safety

AI Safety Summits

- In 2023, the UK held the first AI Safety Summit at Bletchley Park, bringing together governments and AI developers to discuss AI safety for the first time.
- In 2024, the UK co-hosted the AI Seoul Summit. Tech companies, including Google, signed the **Frontier AI Safety Commitments**, setting out an approach for developing frontier AI systems more safely.

The Frontier AI Safety Commitments

The commitments lay out an initial framework for AI developers to take basic precautions against the increasing risks of frontier AI models. They were voluntary commitments, not enforced in any jurisdiction.

Two of the commitments are as follows.

- **"I: Assess the risks posed by their frontier models […] before deploying that model. They should also consider results from […] external evaluations as appropriate."**
- **"VIII: Explain how, if at all, external actors, such as governments […] are involved in the process of assessing the risks of their AI models."**

Google effectively ignored these commitments for their release of Gemini 2.5 Pro, setting a dangerous precedent for future models.

We believe that they should be held accountable for this failure in order to maintain the legitimacy of the international AI Safety Summits and to facilitate the passage of future AI regulation.

Timeline and details of Google's violation

25 March 2025

- *Gemini 2.5 Pro Experimental* becomes available for anyone to access for free. It is arguably the most capable AI model released by any company at this point.
- No information about safety testing is published.

3 April 2025

- Google's head of product for Gemini tells TechCrunch the company hasn't published a 'model card' (ie. safety report) for Gemini 2.5 Pro "because it considers the model to be an 'experimental' release".

9 April 2025

- In correspondence with Fortune magazine, Google does not answer "direct questions" about the involvement of the UK AI Security Institute in the testing process for Gemini 2.5 Pro.
- A spokesperson says they have conducted internal pre-release testing within Google.

16 April 2025

- *Gemini 2.5 Pro Preview* is now available (essentially the same as the *Experimental* model).
- Google publishes its 'model card' (safety report) for Gemini 2.5 Pro Preview.
- The testing report makes no mention of external testing.

28 April 2025

- Google updates its model card to include mention of "third party external testers", but no detail about who they are.

April - June 2025

- Further technical reports have since been published. None of them name the third party external testers or state whether governments have been involved in testing.

Conclusion

- Google violated the spirit of commitment **I** by publishing its first safety report almost a month after public availability and not mentioning external testing in their initial report.
- **Google explicitly violated commitment VIII by not stating whether governments are involved in safety testing, even after being asked directly by reporters.**

General Background on AI

The state of AI

- Tech companies are advancing AI with the explicit goal of building **artificial general intelligence (AGI)**, software that can perform any cognitive task as well as a human.
- **No one knows how quickly future progress will happen.** Google DeepMind CEO, Demis Hassabis, has stated that AGI may be achieved within five years. Other experts say it may be even sooner.
- What is clear is that most **experts were surprised by the extremely rapid progress** in the past five years.

The risks of AI

There are many potential risks from AI, some of which we are already experiencing today. We highlight two that will become increasingly urgent as AI becomes more capable.

1. Economic displacement

- **AGI would be able to perform any job** that a human can do using a computer. With improved robotics, AI could eventually perform almost any work at all.
- This would **concentrate wealth and power** to an unprecedented extent. Countries whose political and economic power does not depend on the well-being of their people, such as those rich in natural resources, are often undemocratic with low standards of living. Similarly, if our economy comes to need few skilled human workers, our own standards of living may decline.

2. Human extinction

- Humans have freedom and sovereignty above other animals due to our intelligence, resourcefulness and coordination. Once AI surpasses us in these traits, we may no longer be able to control our future.
- Hundreds of AI scientists and business leaders have warned us:
  “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”
- In a 2023 survey of over 2,000 top AI researchers, half of respondents estimated a 10% or greater chance that AI would cause human extinction.