Google's violation of Frontier AI Safety Commitments

PauseAI is organising an open letter in response to Google's release of Gemini 2.5 Pro and their violation of their commitments to governments and the public, established in the Frontier AI Safety Commitments.

General Background

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.

Google's Commitments on AI Safety

In 2024, the UK co-hosted the AI Seoul Summit. Google, along with other companies, signed the **Frontier AI Safety Commitments**, setting out initial voluntary commitments for AI safety.

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.".**

Timeline and details of Google's violation

25 March 2025

- *Gemini 2.5 Pro Experimental* becomes available for anyone to access for free.
- No information about safety testing is published.

3 April 2025

- Google's head of product for Gemini tells TechCrunch it hasn't published a 'model card' (safety report) "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.
- A spokesperson says they have conducted internal pre-release testing within Google.

16 April 2025

- Google publishes a bare-bones 'model card' (safety report) for Gemini 2.5 Pro.
- The testing report makes no mention of external testing.

28 April 2025

- Google adds a mention of anonymous "third party external testers" to its model card.

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.**