

AI Security Policy Project

One-Page Briefs on Core AI Security Concepts for French Decision Makers

Summary:

- **Goal:** Producing a series of concise, one-page explainers on foundational AI safety and security concepts, tailored for stakeholders in French institutions.
- **Estimated duration:** 3+ months
- **Requirement:** Contributors must be French-speaking and able to work from our Paris offices.
- **Contact:** [Dr. J  r  my Andr  oletti](#), AI Security Researcher at GPAI Policy Lab

Description

General-Purpose AI (GPAI) systems are increasingly deployed across the economy, with frontier models showing rapid progress in reasoning, autonomy, and task generality. As capabilities grow, so do structural security and control challenges that remain poorly understood outside specialized technical communities. Policymakers often lack concise resources, especially in French, explaining the core mechanisms that can drive systemic AI risks.

This project proposes producing a series of one-page briefs that introduce the main concepts of AI security in a clear, policy-oriented format. All documents will be written in French, and we have a preference for contributors who can come to our Paris office at least a few times.

The series would cover classic topics in the AI Safety and Security literature such as the limits of current interpretability techniques, situational awareness and evaluation awareness, reward hacking, model agency, autonomous replication, etc.

Relevant Expertise

- Strong understanding of AI risks and model behaviour.
- Experience in scientific communication, ability to synthesise complex technical topics.

About the GPAI Policy Lab

The [General-Purpose AI Policy Lab](#) is a non-profit research organization working alongside French (and European) institutions to address the security and international coordination requirements posed by the development of General-Purpose AI (GPAI).

Founded at Sciences Po and based at Campus Cyber in Paris, the GPAI Policy Lab produces research that supports French institutional actors in their efforts to understand, anticipate, and develop concrete solutions to the challenges raised by this technology.

The GPAI Policy Lab aims to serve as a bridge between cutting-edge technical expertise and the operational needs of French security and defense institutions.