

Position: Technical Research and Talent is Needed for Effective AI Governance

Anka Reuel*, Lisa Soder*, Ben Bucknall, Trond Arne Undheim



Overview

Motivation: Past years have seen a lot of governance action on AI. Many of these efforts rely at least to some extent on technical tools and expertise to enact them.

Approach: We surveyed legislation in the EU, US & China to derive areas that need further research for their enactment.

Examples of Gaps in Current Policies



“ Providers of GPAI models with systemic risk shall: perform model evaluation in accordance with standardised protocols and tools

– EU AI Act, Article 55(a)

Open problems: Current evaluations lack robustness, reliability, and validity, especially for foundation models.



“ The Secretary shall require compliance with these [red teaming] reporting requirements for: (i) any model that was trained using a quantity of computing power greater than 1026 FLOP/s

– US Executive Order 14110, Article 4.2

Open problems: Compute thresholds might not be a good measure of risk and we might need other designation criteria



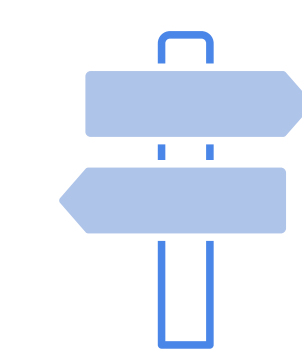
“ Deep synthesis service providers shall employ technical measures to attach symbols to information content produced or edited by their services' users that do not impact users' usage

– Article 7, Provisions on Deep Synthesis Tech.

Open Problems: Current watermarking techniques can be easily spoofed or removed, depending on the modality

The Need for Technical Expertise

➤ **Position:** Work towards a closer integration with policymakers, so as to ensure informed and effective governance of AI.



Inform policy priorities

- Monitoring and communicating key trends in AI development
- Evaluating AI systems to understand current capabilities and impacts



Operationalise policies

- Establishing criteria for the risk classification of AI systems
- Developing guidelines on technical documentation & information sharing



Enforce requirements

- Conducting AI system audits and conformity assessments
- Advising courts on interpreting technical evidence in legal proceedings

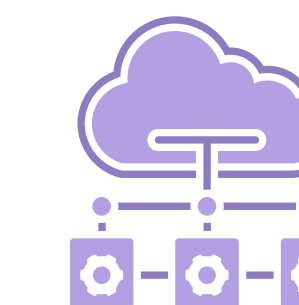
The Need for Technical Research

➤ **Position:** Develop the tools necessary & research that is necessary or can support with enactment of regulatory proposals.



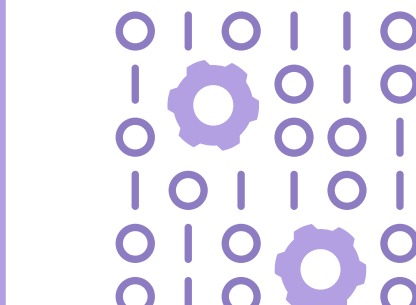
Data

- Identifying sensitive, copyrighted or harmful data in training, fine-tuning, or retrieval datasets
- Detecting or preventing the extraction of training data from AI systems



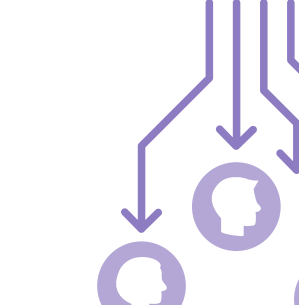
Compute

- Differentiating between AI chip workloads (e.g. training vs. inference) based on chip metadata
- Trusted execution environments on AI chips



Model

- Improving the robustness and reliability of metrics and evaluations of AI systems
- Providing secure researcher and auditor access to AI models



Deployment

- Determining the provenance of AI-generated content
- Evaluating and monitoring the downstream impacts of AI systems

Technical AI Governance (TAIG): We define TAIG as technical tools, research & expertise in support of AI governance. TAIG is only a part of the AI governance toolbox & should be seen in service of sociotechnical & political approaches, rather than as a solution to governance.

Research Agenda: We detail concrete open problems in technical AI governance in a new paper that you can find here

