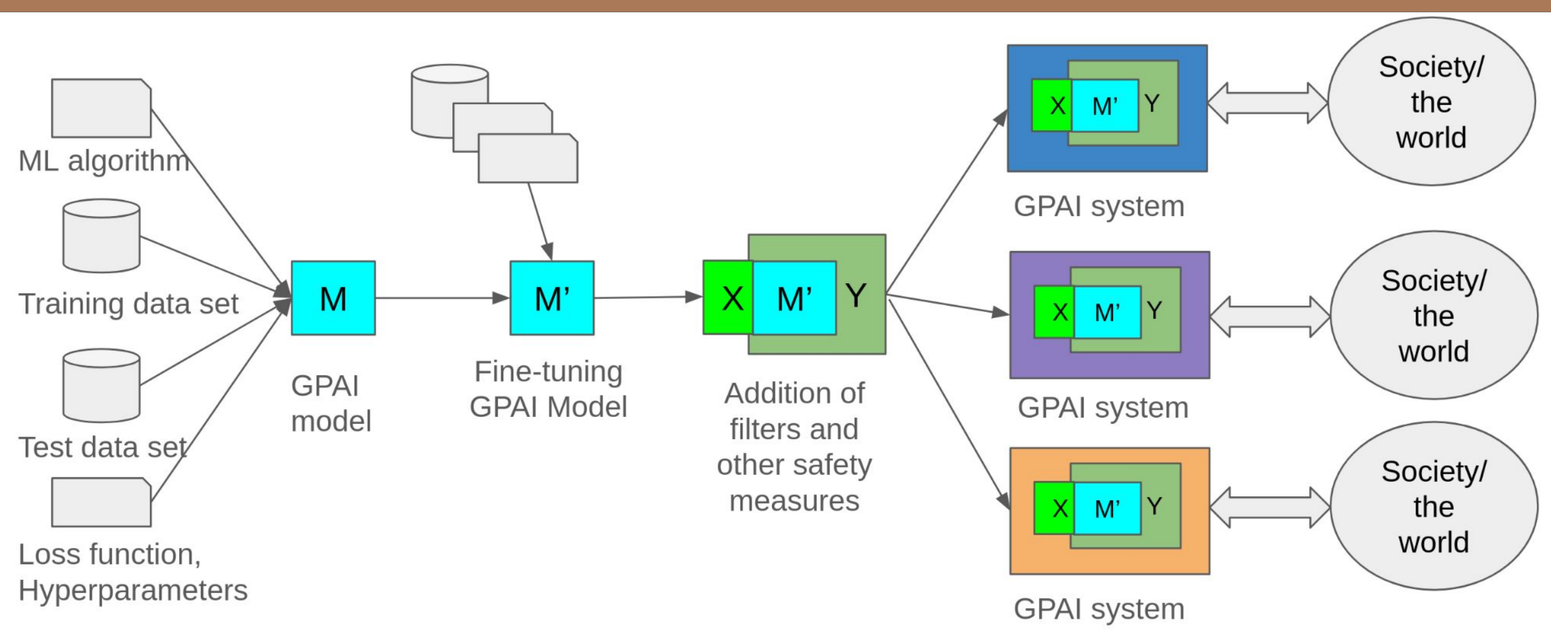
Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems

Rokas Gipiškis^, Ayrton San Joaquin^, Ze Shen Chin*, Adrian Regenfuß*, Ariel Gil*, Koen Holtman*

^ Joint first-authors, *Randomly ordered



Introduction

- ***Key Challenge***
  - Policymakers and companies aim to implement "safe and ethical AI". **But how can they do this in a consistent way? How can we ensure that they were able to do their research?**
- **Standards**
  - The process of defining terms that parties (e.g. companies) need to follow in order for their sector to be consistent.
- **Our Contribution**
  - We create a catalog (A) of the different risks and measures to reduce those risks that stem from any part of the AI model lifecycle. We specifically focus on AI that can perform a wide range of capabilities (GPAI).

Definitions

(1) *Risk* - the combination of the probability of an occurrence of harm and the severity of that harm

(2) Harm: a negative event or negative social development entailing damage or loss to people, property, and the environment

(3) Risk source: an element which alone or in combination has the potential to give rise to risk

(4) Risk management measure: measures that when, applied in its proper context, reduces risk.

(5) Risk Management: The iterative process of identifying risk sources and applying risk management measures to reduce risks to acceptable levels. (B)

Insights and Conclusions

- **Many risk management measures are either in their infancy or require resources that may compete with other interests of a GPAI provider (e.g. financial costs of red teaming)**
  - Regulation can provide implementations of such measures to benefit the worst-off in terms of resources (e.g. individuals or Small-Medium Enterprises).
- **There is a general lack of risk management measures for risk sources in evaluating GPAI models.** GPAI Providers may mistakenly conclude that they have sufficiently managed and evaluated risk given current techniques.

A. Risk Catalog Sample

5.2.1 Benchmark Inaccuracy

Risk source: *Benchmarks may not accurately evaluate capabilities*
Benchmarks of AI systems can both underestimate and overestimate the capabilities of those AI systems.

Underestimates can happen if an evaluation is not comprehensive enough, if the benchmark is saturated by existing models, or if the capabilities in question depend on a complicated setup, such as realistic computer programming tasks.

Overestimates of capabilities can occur if an AI system is trained or fine-tuned on the contents of the benchmark, leading to overfitting.

Risk source: *Benchmark saturation*
Benchmark saturation refers to benchmarks reaching their evaluation ceiling. The tendency towards benchmark saturation has been demonstrated in various benchmarks [19]. When benchmarks reach or are close to saturation, they stop being effective measures for new models, as more nuanced capability gains might not be detected.

Risk management measure: *Statistical data quality reports for benchmarks*
If a benchmark dataset is too large to allow for the identification and removal of all flawed instances, statistical reports on the data composition can be added. Random sampling of benchmark data points can be performed to evaluate and report the frequency and types of errors found [54].

B. Risk Management for GPAI Model Sample