AI SAFETY: AN ENGINEERING PERSPECTIVE

Foutse KHOMH

Professor, Polytechnique Montréal

CIFAR AI Chair, MILA

AI Safety: An Engineering Perspective

Foutse Khomh, Polytechnique Montreal, Mila
foutse.khomh@polymtl.ca

We are witnessing a **shift from AI systems working with "known unknowns"**—planning and learning in uncertain environments to AI systems working in open worlds where **most aspects of the environment are not modeled by the AI agent—the "unknown unknowns".**

To ensure that such AI systems behave safely, we need:

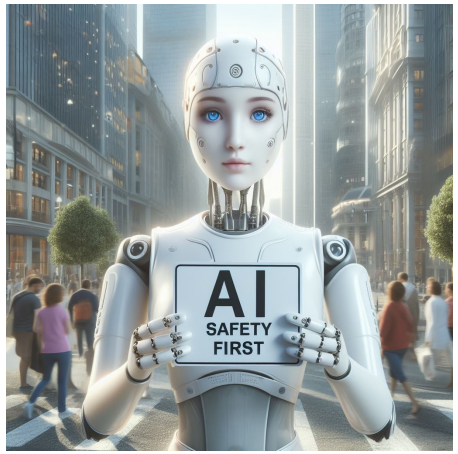
- Novel **principles**, evaluation **methodologies**, and **algorithms** for learning and **acting safely** in the presence of the "unknown unknowns".
- **Robust security protocols** to ensure that AI systems are **safe**, **controllable** and **aligned** with the **intentions** and **instructions** of their designers.

Robust and Resilient AI agents

- How can we ensure that AI agent **behave robustly**, when in an environment different from its training environment?
- How can **we ensure that AI agents do not disturb the environment in negative ways** while pursuing their goals?

Aligned AI agents

- How can we ensure that **an AI agent won't game its reward function**?
- How can we efficiently **ensure that an AI agent respects aspects of the objective** that are too expensive to be frequently evaluated during training?



Regulated AI agents

- How can we design policies, regulations and guidelines **to govern the safe development, maintenance, and deployment** of AI agents?
- How can we **effectively** and **efficiently continuously**, **verify**, **monitor**, and **enforce policies and regulations for safe and secure** cooperation between humans, AI agents, and other systems?

Mila
SEMLA
Trustworthy Engineering of AI Software
CIFAR
POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE