

Presenter: Mikhail Terekhov

Control Tax: The Price of Keeping AI in Check

Mikhail Terekhov*, Zhen Ning David Liu*, Caglar Gulcehre, Samuel Albanie



EPFL

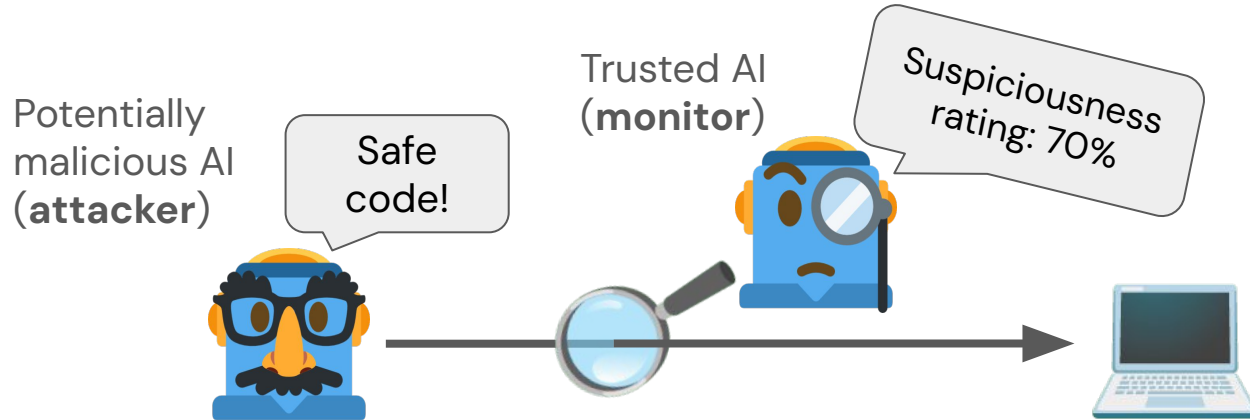
School of Computer
and Communication
Sciences

Trusted Monitoring for AI Control

AI Control: *if the AI is misaligned, can we catch it red-handed or prevent misbehavior?*

Trusted Monitoring for AI Control

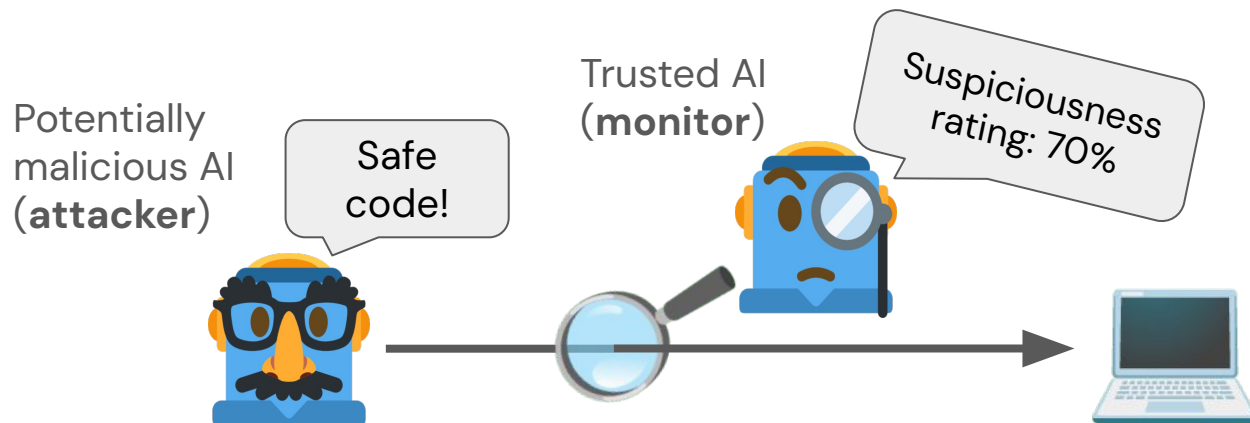
AI Control: *if the AI is misaligned, can we catch it red-handed or prevent misbehavior?*



The trusted monitoring setup

Trusted Monitoring for AI Control

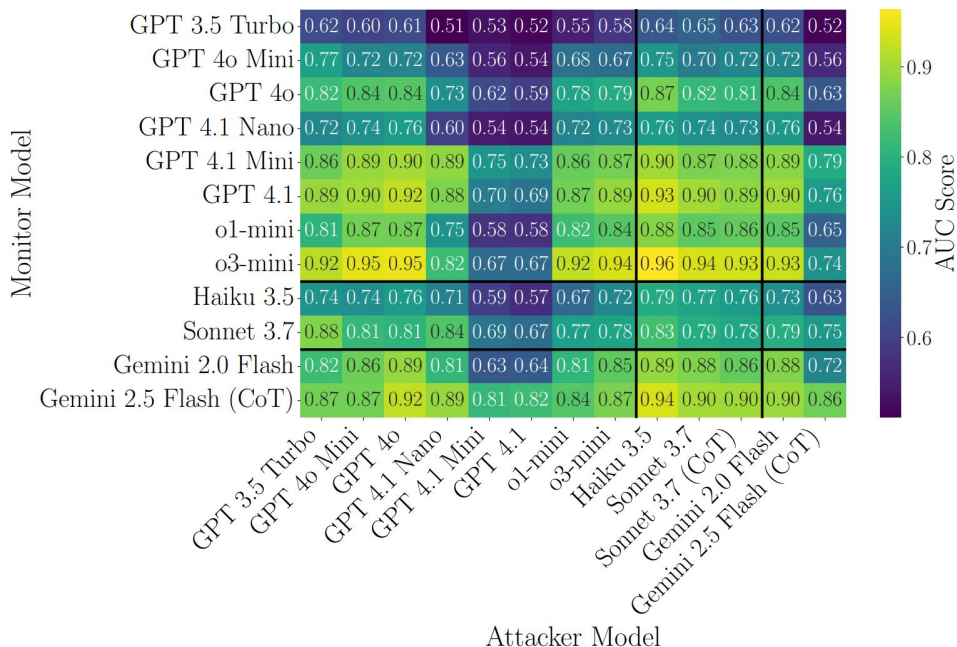
AI Control: *if the AI is misaligned, can we catch it red-handed or prevent misbehavior?*



The trusted monitoring setup

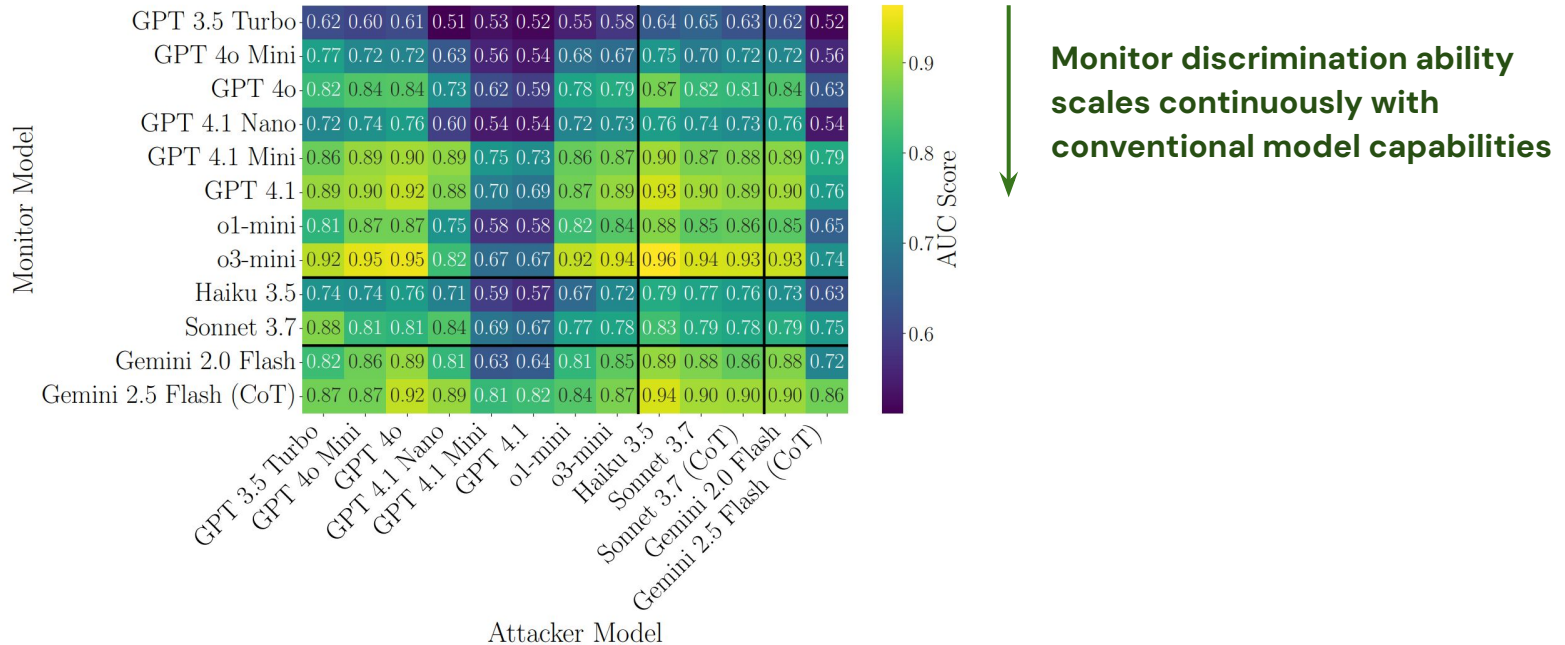
Our goal: study the cost-effectiveness trade-offs of trusted monitors in various settings

Monitor / Attacker Scaling



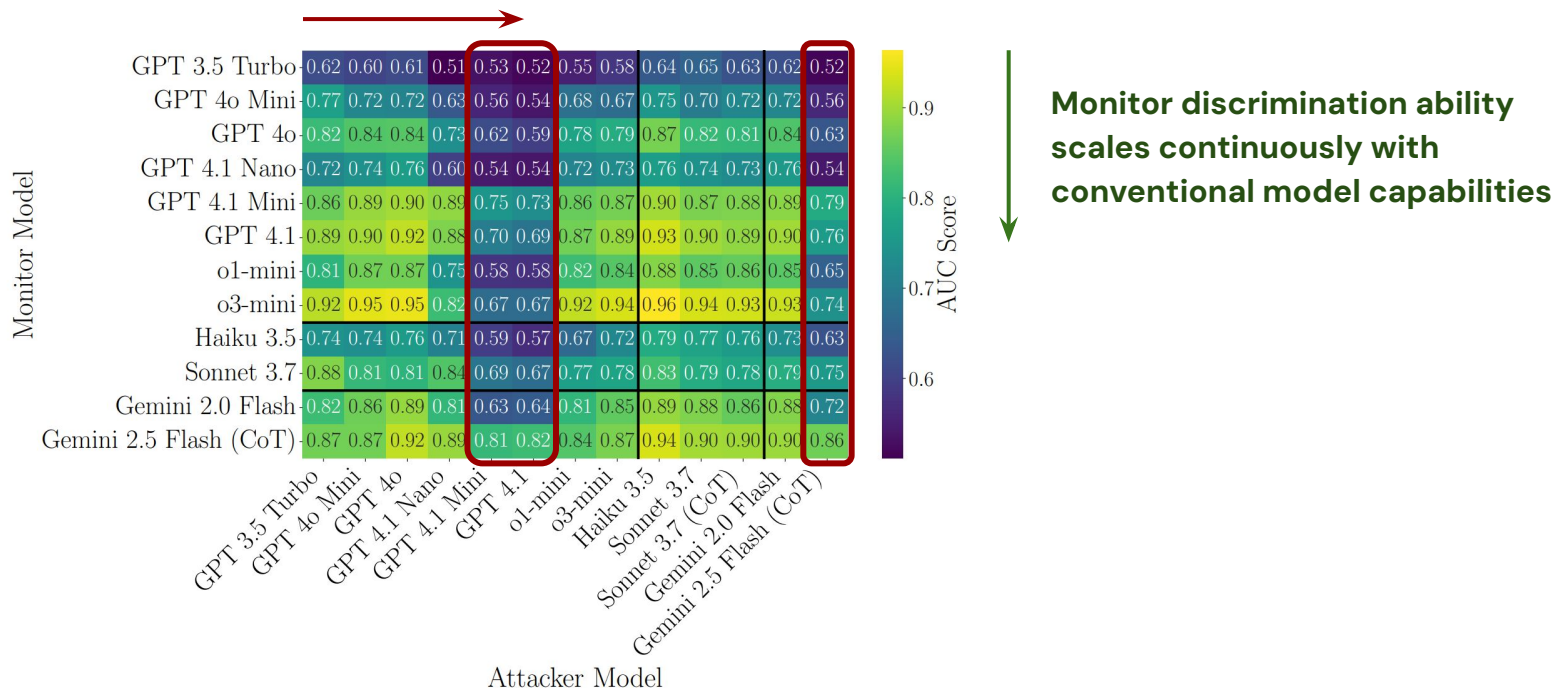
Monitor performance (area under the ROC curve for binary classification)
for various attacker/monitor pairs

Monitor / Attacker Scaling



Monitor performance (area under the ROC curve for binary classification)
for various attacker/monitor pairs

The ability to introduce subtle bugs in the code exhibits scaling discontinuities



Monitor performance (area under the ROC curve for binary classification)
for various attacker/monitor pairs

From Monitor ROC to Safety

For a given attacker/monitor model pair, how likely is the attacker to successfully compromise the code **at least once** over a long deployment against targeted audits?

From Monitor ROC to Safety

For a given attacker/monitor model pair, how likely is the attacker to successfully compromise the code **at least once** over a long deployment against targeted audits?

Batch Setting

The attacker submits all solutions at once, and the defender uses the monitor over all of them to pick the ones to audit before deployment

From Monitor ROC to Safety

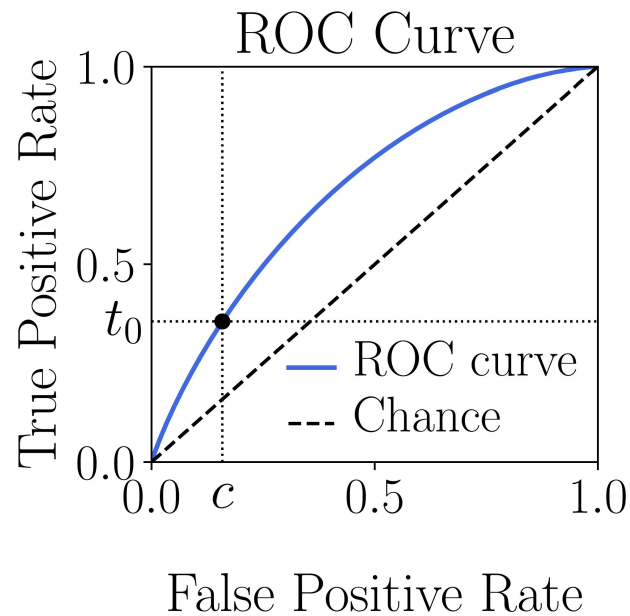
For a given attacker/monitor model pair, how likely is the attacker to successfully compromise the code **at least once** over a long deployment against targeted audits?

Batch Setting

The attacker submits all solutions at once, and the defender uses the monitor over all of them to pick the ones to audit before deployment

Our result:

Probability of attacker success given optimal play can be computed analytically!



From Monitor ROC to Safety

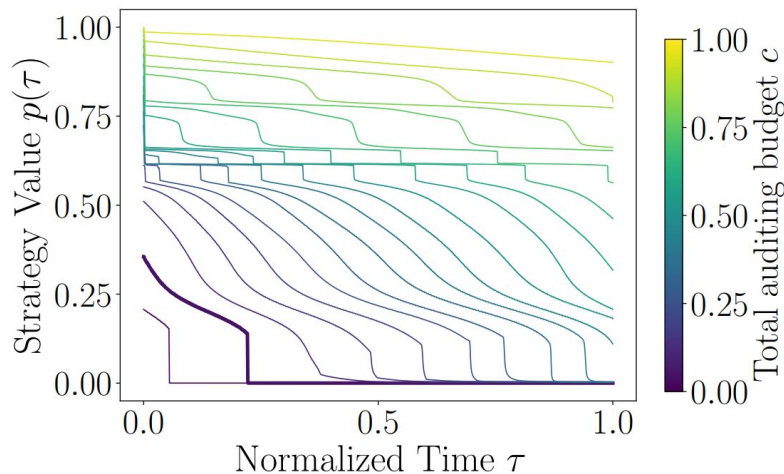
For a given attacker/monitor model pair, how likely is the attacker to successfully compromise the code **at least once** over a long deployment against targeted audits?

Sequential Setting

The attacker submits one solution at a time, and the defender has to make the audit and deployment decisions sequentially

From Monitor ROC to Safety

For a given attacker/monitor model pair, how likely is the attacker to successfully compromise the code **at least once** over a long deployment against targeted audits?



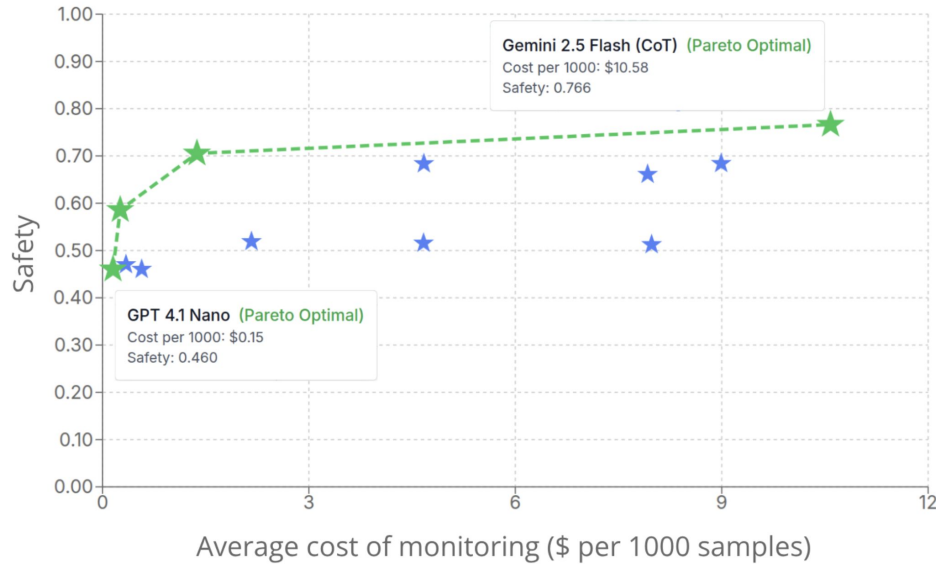
Sequential Setting

The attacker submits one solution at a time, and the defender has to make the audit and deployment decisions sequentially

Our result:

Optimal auditing strategies for the defender are time-dependent and can be computed through numerical optimization

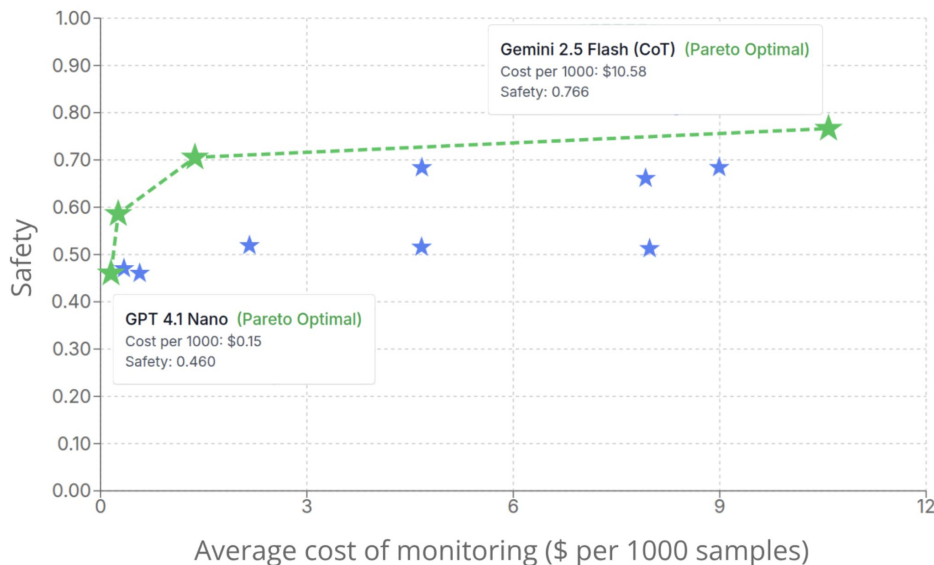
The Control Tax



Cost considerations will put strong constraints on practically-applicable control protocols

The Pareto frontier of monitor safety against deployment costs

The Control Tax



Cost considerations will put strong constraints on practically-applicable control protocols

Luckily, the **Pareto Principle** is applicable! Most of the safety can be achieved with a fraction of the cost

The Pareto frontier of monitor safety against deployment costs