Haojian Jin

Where are we in the class?

- System
  - Location privacy; Permissions for Privacy; Policies for Privacy
- User
  - Privacy Norms/Contextual Integrity
  - Individual Privacy (CogSci); User Agency
- Developer | Auditors
  - Privacy designs
  - Privacy regulations
- AI Safety/Privacy

Agenda

- EU AI Act
- AI Fairness/Ethics
- GenAI Privacy

EU AI Acts (2021-2023)

Unacceptable Risk

Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

High Risk

Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Limited Risk

Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

Minimal Risk

All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

EU AI Acts - Prohibited AI Practices

- Placing on the market of an AI system that causes physical or psychological harm of a person
- General purpose social scoring by public authorities
  - Extension to private actors.
- Use of real-time remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement
  - Exemptions? Law enforcements?

EU AI Acts - Prohibited AI Practices (Continue)

- Categorise people according to sensitive attributes.
- Assesses the risk of a person (or group of persons) to commit criminal or administrative offenses based on profiling or assessing personality traits and characteristics.
- Facial recognition from internet or CCTV footage
- Infer emotions in the categories of law enforcement, border management, in workplace and education institutions

Reflection

Amazon's Ring is the largest civilian surveillance network the US has ever seen

Lauren Bridges

One in 10 US police departments can now access videos from millions of privately owned home security cameras without a warrant

EU AI Acts - High-risk systems

AI system which do not lead to a significant risk to the health, safety or fundamental rights should not be considered as high-risk.

EU AI Acts - High-risk AI Systems

- Biometric identification of people
- Safety components in the management of critical infrastructure (road traffic, water, gas, heating & electricity)
- Education and vocational training
- Employment, workers management and access to self employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management

EU AI Acts - High-risk AI Systems - Continued

- Administration of justice and democratic processes
- AI used as safety component of certain products (incl. personal protective equipment & medical devices)
- AI in the evaluation and classification of emergency calls
- AI to be used as safety components in the management and operational of critical digital infrastructure
- AI to be used for influencing the outcome of an election of referendum
- AI to be used by large social media platforms to recommend user-generated content available on the platform

EU AI Acts - High-risk AI systems - Example

- Access to and enjoyment of essential private services and public services and benefits
  - "AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid”

EU AI Acts - High-risk AI systems - Example

- Law enforcement
  - Assessing the risk of people for committing criminal offences.
  - Detecting the emotional state of a person
  - Detecting deep fakes (later removed.)
  - Evaluating the reliability of an evidence
  - Predicting the occurrence of a criminal offence based on profiling people
  - Profiling people in the course of detectoin, investigation or prosecution.
  - Crime analytics to search complex large data for investigation purposes

How to work with High Risk AI Systems

- Risk management system
- Data and data governance
- Technical documentation
- Record-keeping
- Transparency and provision of information to users
- Human oversight
- Accuracy, robustness and cybersecurity

Specific obligations

- Providers
  - Quality Management System
  - Draw up technical documentation
  - Conformity Assessment
  - Keep the logs automatically generated
  - Cooperation with competent authorities
  - Appointment of legal representative
- Importers
- Distributors
- Users

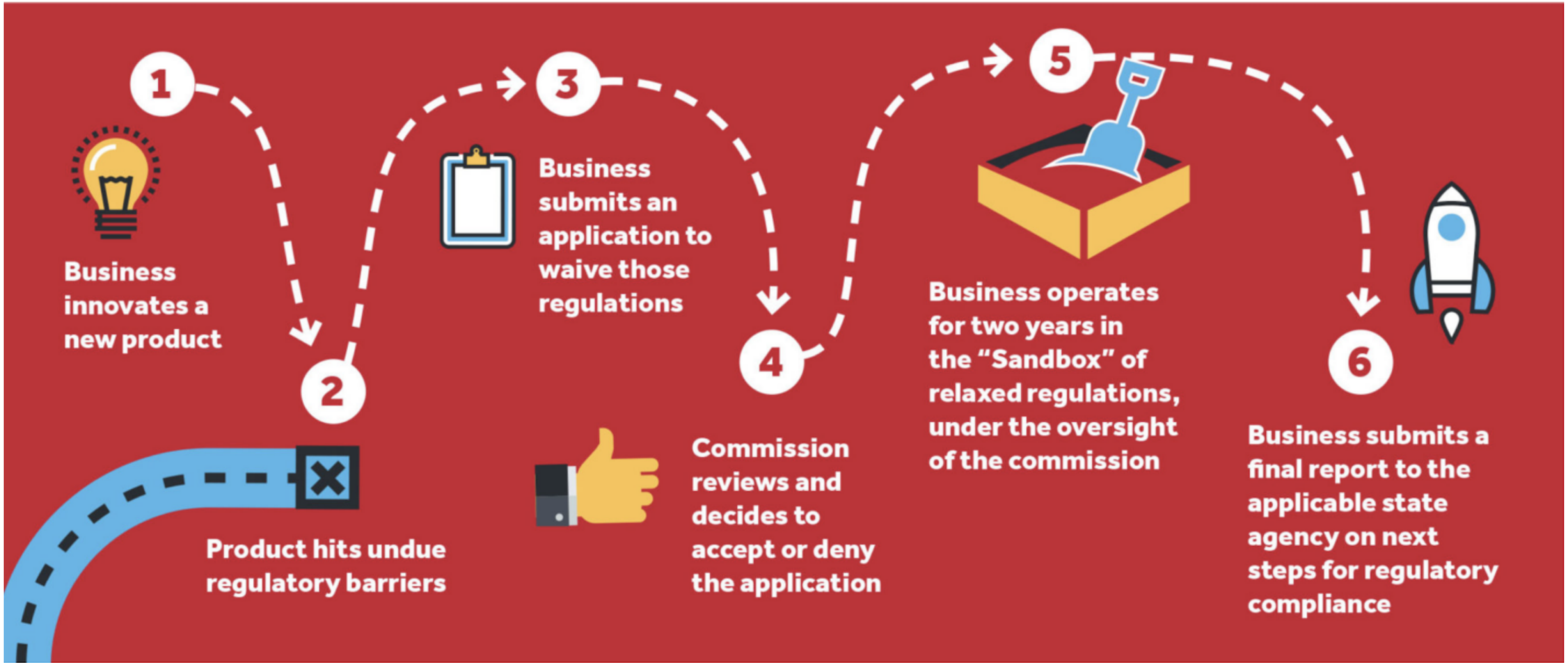
Specific obligations

- Providers
- Importers
- Distributors
- Users
  - Monitor the operation of the AI system
  - Keep the logs automatically generated
- Amendment: Provider of a foundation model

High Risk AI Systems - EU Database

- European Commission to set up a database for stand-alone high-risk AI systems.
- Mandatory registration
- Should be accessible to the public.

High-risk AI Systems - regulatory sandboxes



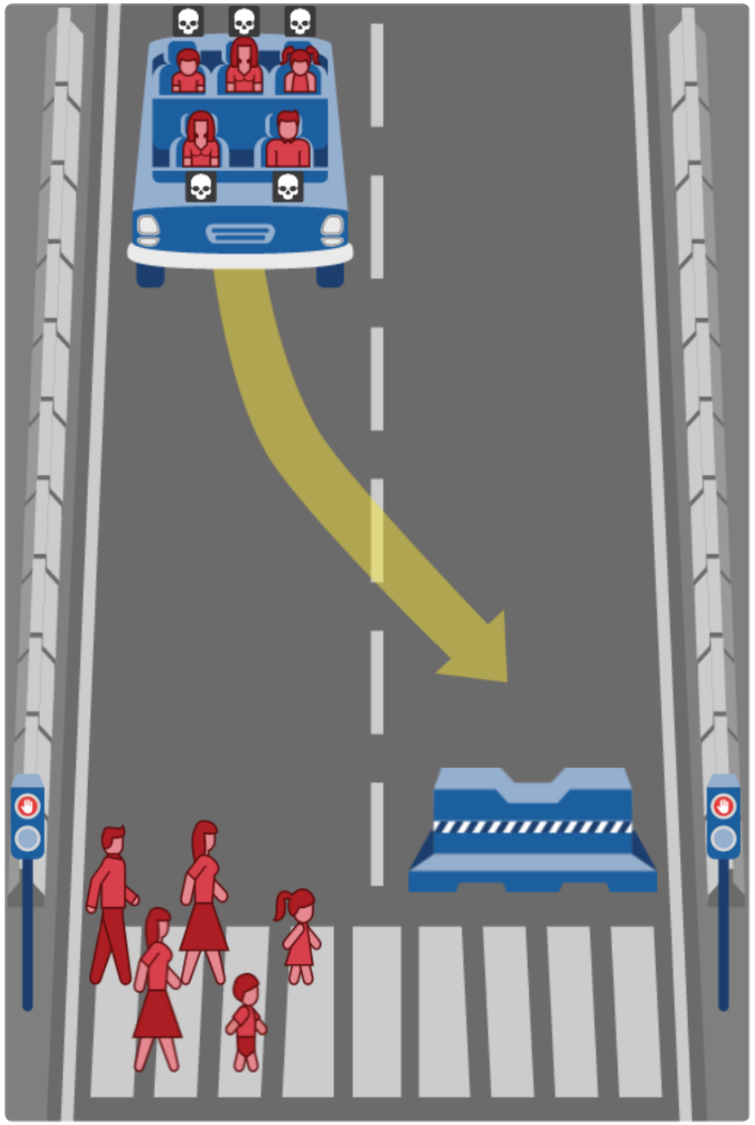
EU AI Acts - Sum up

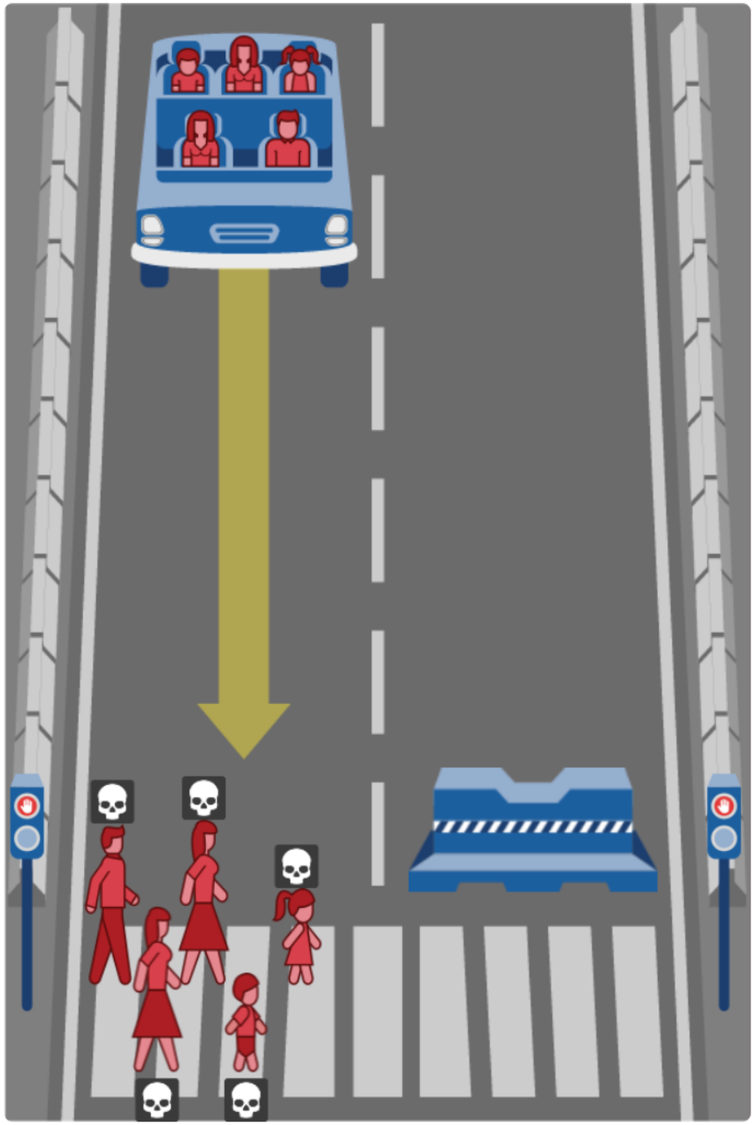
- AI will not be banned.
- Some specific practices will be prohibited.
- High-risk AI systems should come with specific protocols, documentation, strategies…

Agenda

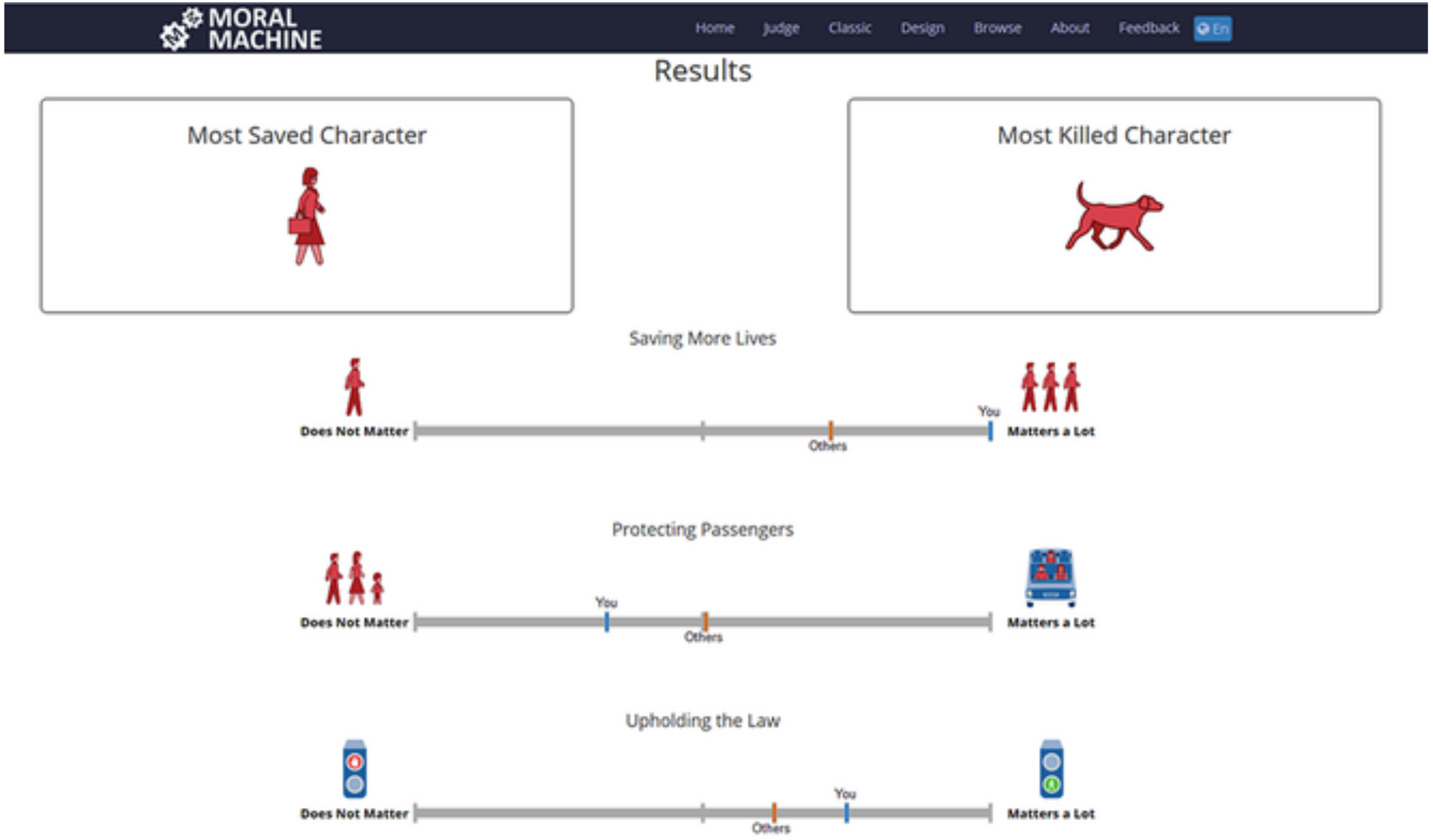
- EU AI Act
- AI Fairness/Ethics
- GenAI Privacy

Moral Machines





Moral Machine



arXiv

https://arxiv.org › cs

[1610.02413] Equality of Opportunity in Supervised Learning

by M Hardt · 2016 · Cited by 5424 — We propose a criterion for discrimination against a specified sensitive attribute in **supervised learning**, where the goal is to predict some target based on ...

Demo: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Agenda

- EU AI Act
- AI Fairness/Ethics
- GenAI Privacy

Primer on AI Alignment

- Humans specify what we want from ML systems via
  - Human feedback (rewards)
  - Natural language instructions
- Definition of harm not reducible to simple formulae
- AI misalignment = mismatch between AI behavior and human intentions
- Many documented examples with LLMs, including confidently wrong statements (hallucinations)

Primer on AI Safety

- Technology is dual-use
- Intellectual capability + bad goals = danger
- Are we individually and collectively wise enough?
- Can we lose control of a human-designed AI system?
- How do we minimize those risks?

Alignment challenge

- Making sure that AI behaves according to our intentions
- Makes it difficult to design AIs that behave morally

- **Unsolved challenge** up to now – *and it is not even clear that a solution exists*

- **Potential danger of loss of control of AI**
  - *AI's self-preservation objective*
    - Emerges as a side-effect of misalignment/reward hacking
    - Provided by misguided humans, "Frankenstein scenario"

→ new potentially dangerous species could dominate humans

Why the concern about dangerous AI

- Brains = biological machines

- Computers have advantages over brains!
  - Faster learner
    - Backprop, precise memory, communication bandwidth
  - Human-level intelligence in a few years maybe decades?

- Computers/software are accessible!
  - Rogue or misguided humans could create powerful & dangerous AI

Democracy vs Power Concentration

- AI in a few hands = power concentration
- Increased AI capability = increased power concentration
- Superhuman AI in a few (public or private) hands = extreme power concentration = capacity to profoundly influence, destabilize or dictate political, social, miltary and economic agendas

→ **Future Frontier AIs can threaten social order and markets**

EVEN IF WE SOLVE SAFETY/ALIGNMENT

- **Governments must acquire the capability to master AI**

What we can do to mitigate harms & risks?

3 recommendations in testimony to US senate (July 2023):

1. National **regulation** + international treaties, reducing access and frontier development to licensed organizations & registered models, with audits, banning models not proven safe
2. Major research investment in **AI safety**, to better understand risks and guide regulation (what threshold for danger? How to evaluate and benchmark risks?)
3. Plan B against the eventual emergence of AI-driven threats from misuse or loss of control (**countermeasures**)

How to build a powerful safe AI?

- What is **not safe**?

Superhuman capabilities

+

Agency

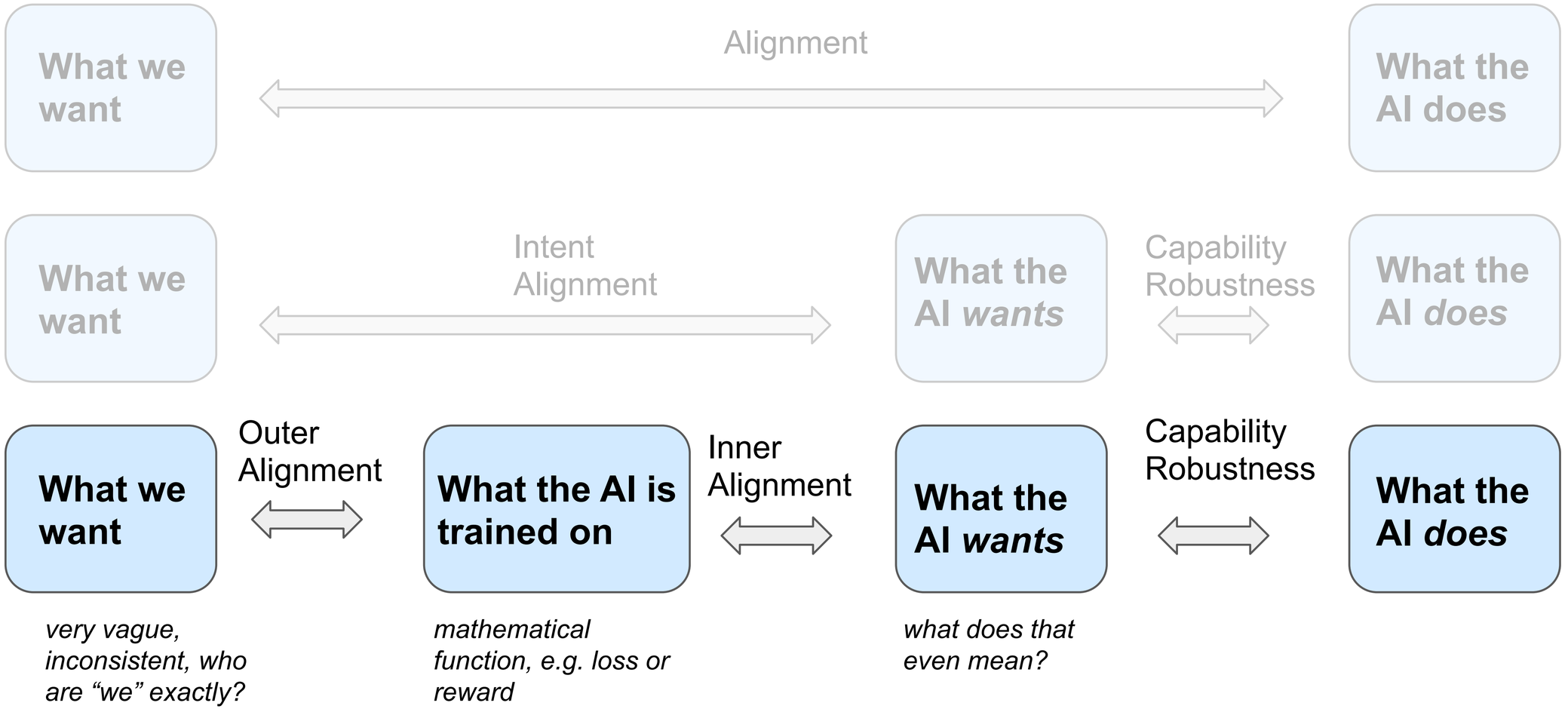
+

Misalignment & its amplification

AGI might kill us!

Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.

The gaps



(Very simplified. Doesn't account for multipolar AI. Also a lot of alignment research doesn't neatly fit into these categories. It’s also disputed if the inner/outer distinction is productive)