Regulating AI?

Challenges and opportunities for the development of safe and ethical AI

Dr Carina Prunkl
Future of Humanity Institute, University of Oxford
carina.prunkl@philosophy.ox.ac.uk
Risks from AI

◊ Accidents
Risks from AI

- Accidents
  - Malfunctioning
Risks from AI

✧ Accidents
  ✧ Malfunctioning
  ✧ Unintended consequences
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use

Carina Prunkl, University of Oxford, carina.prunkl@philosophy.ox.ac.uk
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use
- Inadequate use
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use
- Inadequate use
- Structural changes
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use
- Inadequate use
- Structural changes
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use
- Inadequate use
- Structural changes

Contributing Factors
Risks from AI

- Accidents
  - Malfunctioning
  - Unintended consequences
- Malicious use
- Inadequate use
- Structural changes

Contributing Factors

- AI capabilities
- Access and scale of deployment
- Speed of deployment
Regulatory challenges

- Risk uncertainty
- Context dependence
- Speed of development
  - Regulatory inertia
- Incremental changes
- Experts
  - Scarcity of talent
  - Regulatory capture
- Globalisation
Self-regulation?

Solves some existing problems:
- Expertise
- Efficiency

Current attempts:
- Standards
- Coalitions (e.g. Partnership on AI)
- Principles and codes
AI Principles

- More than 100 AI principles by governments, companies, NGOs, professional bodies, research institutions, etc.

- Arbitrariness

Source: https://ai-hr.cyber.harvard.edu/primp-viz.html
AI Principles: the value landscape

- **Fairness**
  - Does the technology enable or contribute to individual or structural unfairness?

- **Autonomy**
  - Does the use of the technology interfere with human autonomy or freedom, i.e. through manipulation, deception or coercion?

- **Safety**
  - Are there actual or potential physical, mental or structural harms associated with the use of the technology?

- **Privacy**
  - How is personal data collected, stored, used and distributed?

- **Transparency**
  - When, how and what technology is used by whom and for what purpose?

- **Explainability**
  - Can we understand how a decision was made?

- **Accountability**
  - Who is responsible if something goes wrong?
AI Principles

- More than 100 AI principles by governments, companies, NGOs, professional bodies, research institutions, etc.

- Arbitrariness
  - Why these, not other values/principles?

Source: https://ai-hr.cyber.harvard.edu/primp-viz.html
AI Principles

- More than 100 AI principles by governments, companies, NGOs, professional bodies, research institutions, etc.

- Arbitrariness
  - Why these, not other values/principles?

- Lack of guidance
  - How to act when values conflict?

Source: https://ai-hr.cyber.harvard.edu/primp-viz.html
Principles are not enough

- Arbitrariness, vagueness and lack of guidance make them insufficient governance tools

- Implementation of these principles into governance solutions
  - But too context specific?
  - Conflicting incentive structures lead to suboptimal implementation
  - But only in some cases

- Alternative forms of governance?
  - Public regulation
  - Standards
  - ML research community
Conclusion

- Risk factors from AI:
  - Accidents
  - malicious use
  - inadequate use
  - structural risks

- Governance mechanisms:
  - Legal and regulatory: risk uncertainty
  - Self-governance: Principles
    - Vague, unhelpful, often merely signalling

- Alternative forms of governance:
  - Standards
  - ML research community

Thank you for your attention.

@carinaprunkl
carina.prunkl@philosophy.ox.ac.uk