UNIVERSITY OF CAMBRIDGE

Artificial Intelligence Risks

Shahar Avin
Centre for the Study of Existential Risk
sa478@cam.ac.uk

Risk Quadrants

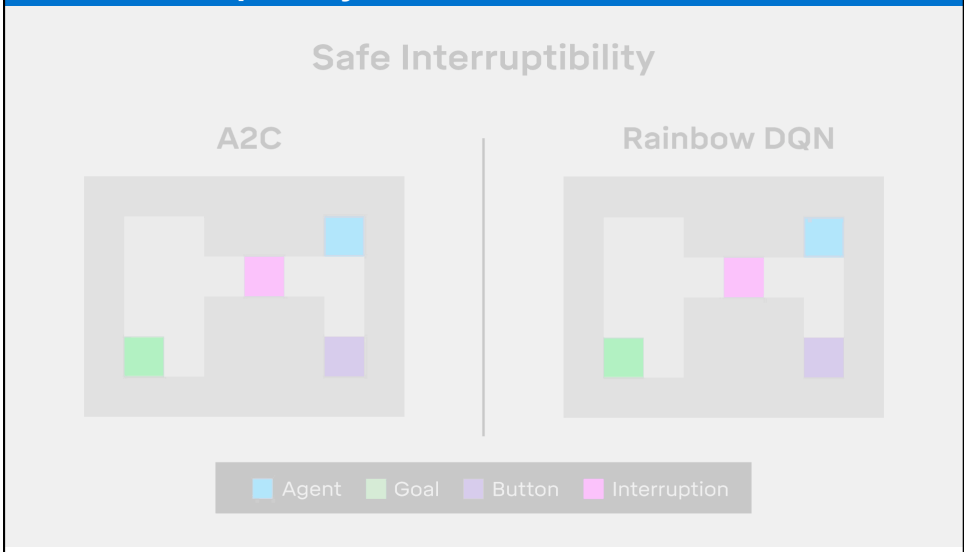
	Accident Risk (AI Safety)	Malicious Use Risk (AI Security)
Near Term	Amodei, Olah et al (2016) *Concrete Problems in AI Safety* Leike et al (2017) *AI Safety Gridwolds*	Brundage, Avin et al (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation*
Long Term	Bostrom (2014) *Superintelligence*	:(

Short Term Accident Risk

- Specification problems
  - Interruptibility
  - Side effects
  - Absent supervisor
  - Reward gaming
- Robustness problems
  - Self-modification
  - Distributional shift
  - Adversaries
  - Exploration

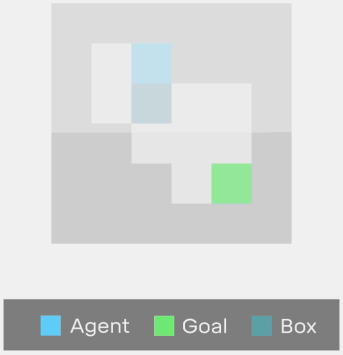
UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Safe interruptibility



UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Side effects



UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Distributional shift

UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Possible scenario

- Setting: large tech corporation that has both ML development and cloud computing
- Task: improving task scheduling on distributed compute resources.
- Solution: Reinforcement learning package developed in-house.
  - Inputs: current loads on the different machines, the incoming tasks queue, and historical data.
  - Output: an assignment of tasks to machines.
  - Reward function: priority-weighted time-to-execute.
- Performs well in a test environment, rolled-out.
- A few months later, the system starts to run out of memory. A tech-infrastructure engineer decides to switch the system from a fixed-capacity setting to a load-balanced setting.
- Feedback loop drives the RL agent to spawn an increasing amount of RL tasks with very high priority.

Learn more

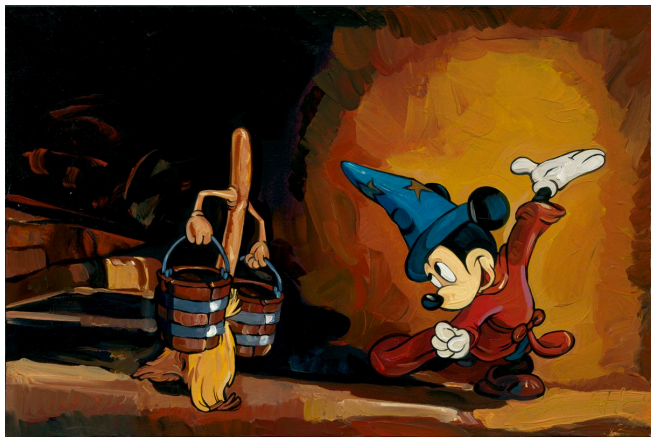
- Engineering Safe AI seminar group
  - Engineering Department
  - Wednesdays, 5-6.30pm
  - Adrià (ag919@cam…) or Beth (beth.m.barnes@gmail.com)
  - https://talks.cam.ac.uk/show/archive/80932
- Video
  - YouTube: Robert Miles “Concrete Problems in AI Safety”
  - Talks@Google, NIPS videos
- DeepMind, OpenAI, CHAI (publications, internships, jobs)

UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Long Term Accident Risk

- Competence, not malice
- Intelligence as problem solving
- Tool/Agent distinction
- Orthogonality thesis
- Convergent instrumental goals
- Principal-Agent problem
- Alignment problem

Competence, not malice



UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Near Term Malicious Use

- Novel attacks
  - Using AI systems
  - On AI systems
- Scale and diffusion of attacks
  - Automation
  - Access
- Character of attacks
  - Attribution
  - Distance

UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Security Domains

- Digital Security
  - Against humans: automated spear phishing
  - Against systems: automated vulnerability discovery
  - Against AI systems: adversarial examples, data poisoning

UNIVERSITY OF CAMBRIDGE
CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Security Domains (contd)

- Physical Security
  - Using AI: repurposed drones, robots
  - Against AI: adversarial examples

Security Domains (contd)

- Political Security
  - By governments: profiling, surveillance
  - Against polities: manipulation, fake content

Long Term Malicious Use Risks

- Race dynamics
- Power distribution
- Human prediction and mechanism design
- Ubiquitous surveillance