PURDUE UNIVERSITY | Department of Computer Science

CS57100: Artificial Intelligence
Ethics and AI

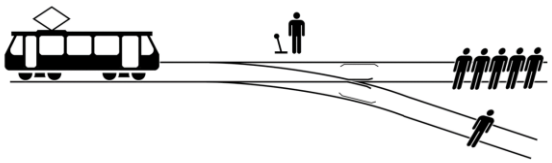
Prof. Chris Clifton

9 November 2022

Indiana Center for Database Systems™

PURDUE UNIVERSITY
Department of Computer Science

The Trolley Problem



- Should you pull the lever?
- How do we encode this in AI?
- Current variant: Autonomous vehicles
  - Do I kill my occupant or a pedestrian?
  - *Today's AI won't be able to ASK the question*
- Better AI response:
  Stop the train before getting to this point!

2

PURDUE UNIVERSITY
Department of Computer Science

Outline

- Use Cases
  - Autonomous weapons
  - Impact on people
- Limits of AI
  - Safety
- Decisions
  - Trolley problem
  - Discrimination
- Privacy
- Trust/Transparency
- Rights of AI
  - Legal personhood?
  - Intellectual Property?
- Ethical Reasoning
  - History

3

PURDUE UNIVERSITY
Department of Computer Science

AI Use Cases: *Autonomous Weapons*

- Improvements in weaponry can reduce loss of life
  - Incapacitate / "neutralize" without lethality
  - Precision-guided munitions, avoid collateral damage
  - Reduce risk to troops/police
- Do we want AI to make a call intended to (or highly likely to) cause loss of life?
  - Do we shoot the unidentified (missile? airliner? eagle?) before it hits the building?
  - *You have 20 milliseconds to answer or it will be too late…*

4

PURDUE UNIVERSITY
Department of Computer Science

AI Use Cases: *Lethal* Autonomous Weapons

- DODD 3000.09: weapon system[s] that, once activated, can select and engage targets without further intervention by a human operator
  - Does this include a land mine?
- Requires weapons systems be designed to "allow commanders and operators to exercise appropriate levels of human judgement over the use of force."
  - Not necessarily human control, but determination of how/when/where/why used

5

PURDUE UNIVERSITY
Department of Computer Science

AI Use Cases: *Lethal* Autonomous Weapons

- UN Convention on Certain Conventional Weapons
  - Started discussions in 2014
  - Status "upgraded" in 2018
- Adopted 11 guiding principles in 2019

a) International law applies
b) Human responsibility must remain
c) Human-machine interaction must ensure compliance with international law
d) Human accountability at all stages
e) States must ensure compliance with international law
f) Risk of acquisition by terrorist groups must be considered
g) Risk assessment and mitigation should be part of development
h) Comply with IHL and other international legal obligations
i) Do not anthropomorphize LAWS
j) Policy should not prevent peaceful use
k) Exiting CCW provides appropriate framework

6

PURDUE UNIVERSITY
Department of Computer Science

Top Ethical Issues
As presented at 2016 WEF

1. Unemployment
2. Distribution of machine-created wealth
3. Impact on human behavior/interaction
4. Guarding against mistakes
5. AI bias
6. Safety from adversaries
7. Protect against unintended consequences
8. How do we stay in control?
9. Robot rights

7

PURDUE UNIVERSITY
Department of Computer Science

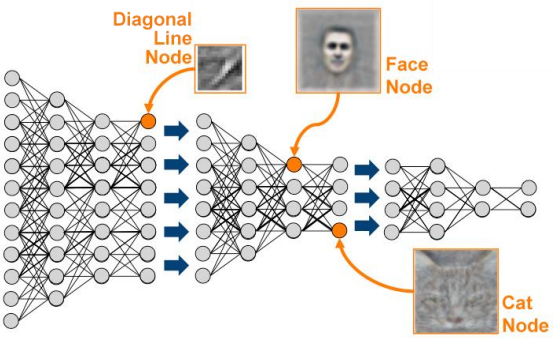
Ethical Issues:
AI Safety

- Multiple issues
  - Mistakes
  - Unintended consequences
  - Protection from adversaries
- Can we guarantee certain outcomes?
  - Rule out bad outcomes?

8

PURDUE UNIVERSITY
Department of Computer Science

How Does it Work? Image Recognition with Deep Learning



Source: "The Web" (many sources, none credit the original creator)

- Understand?
- What AI "sees" and what we see may be completely different

10

PURDUE UNIVERSITY
Department of Computer Science

AI Safety: Solutions

- External controls
  - "kill switch"
- Constraints on AI
  - Prove that certain properties will be satisfied
  - *Program Verification*
- Randomization
  - "Moving Target" defense
  - Ensure variability in AI decision-making

11

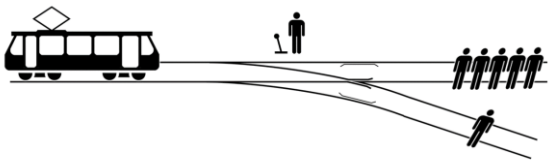
PURDUE UNIVERSITY
Department of Computer Science

Outline

- Use Cases
  - Autonomous weapons
  - Impact on people
- Limits of AI
  - Safety
- Decisions
  - Trolley problem
  - Discrimination
- Privacy
- Trust/Transparency
- Rights of AI
  - Legal personhood?
  - Intellectual Property?
- Ethical Reasoning
  - History

PURDUE UNIVERSITY
Department of Computer Science

The Trolley Problem



- Asimov's First Law of Robotics
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- ???