Robots, Ethics, and Robot Ethics

willie

Show ‘n Tell

Meet Jibo

Why robots?

Interacting with a real environment

Ethics in AI

Automation and the workplace

Anthropomorphism

Data Bias, Algorithmic fairness

Explainability and interpretability

Moral responsibility, Blameworthiness

Privacy

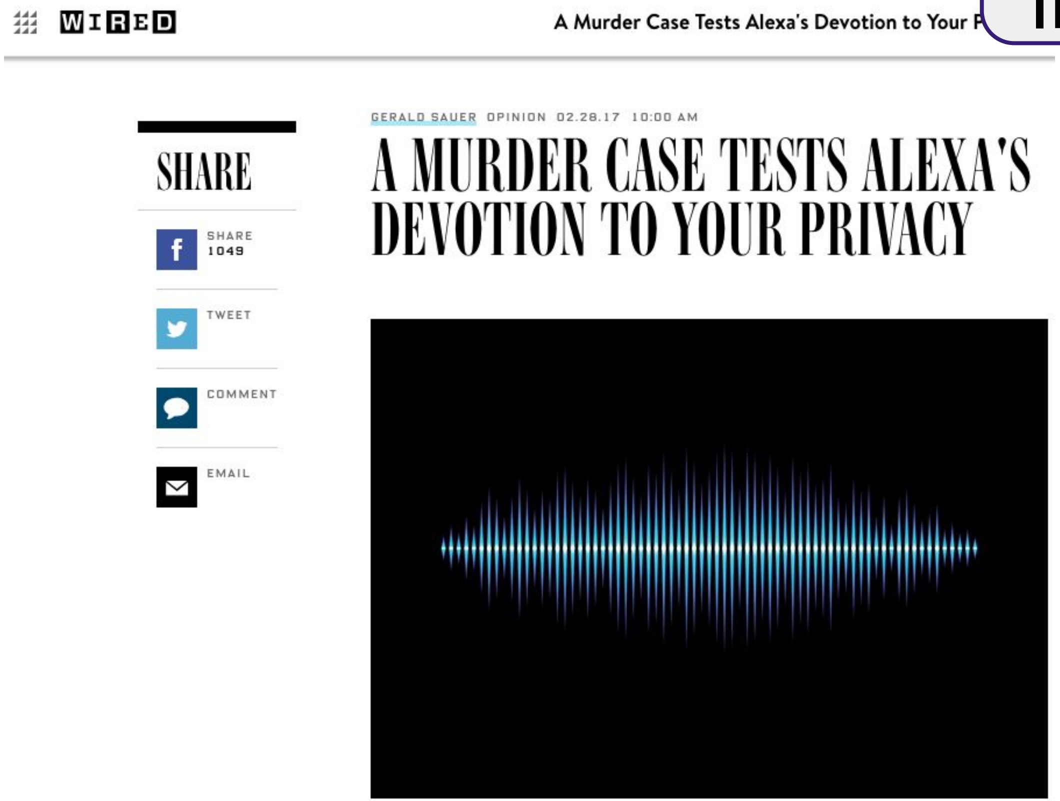
Weaponization

Robots rights

Unintended consequences

Robot obedience

Sensors are everywhere!



WIRED

A Murder Case Tests Alexa's Devotion to Your P

SHARE

SHARE 1049

TWEET

COMMENT

EMAIL

GERALD SAUER OPINION 02.28.17 10:00 AM

A MURDER CASE TESTS ALEXA'S DEVOTION TO YOUR PRIVACY

https://www.wired.com/2017/02/murder-case-tests-alexas-devotion-privacy/

* Sort of. It is always listening for the activation word, but supposedly nothing more.

Bias 1

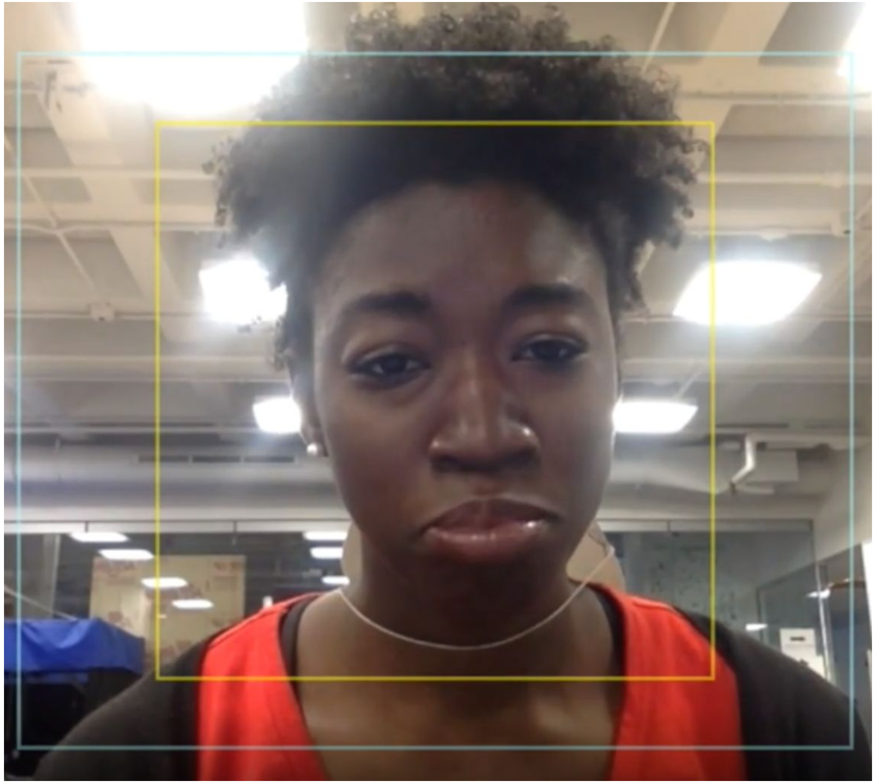
Algorithmic bias

Commercial face detection from:

- Microsoft
- IBM
- Megvii

Error rates

- 21-35% dark-skinned women
- Below 1% light-skinned females



https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Bias 2

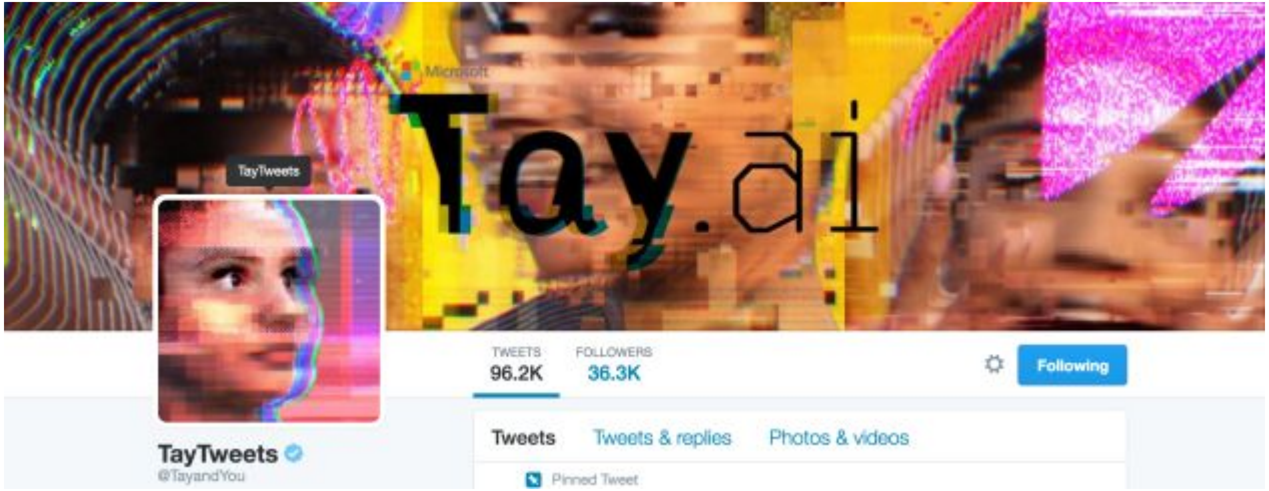
Millennial-minded AI

Casual conversation

Design intent: more it’s used, the better it gets

Mimicry

What could possibly go wrong?



https://www.geekwire.com/2016/even-robot-teens-impressionable-microsofts-tay-ai-pulled-internet-teaches-racism/
Warning: includes content that may be highly offensive

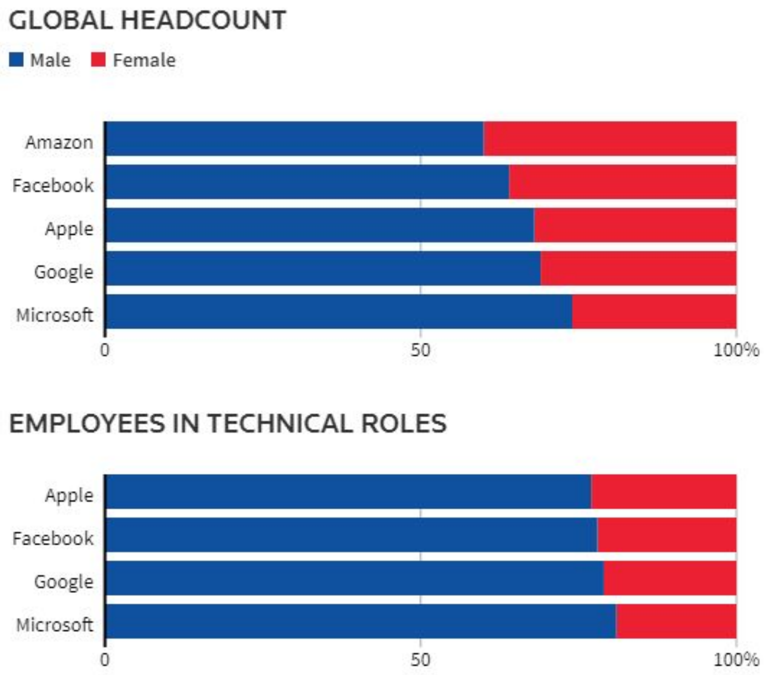
Bias 3

Amazon built a system to more quickly find the top applicants from a pool

Trained on 10 years of resumes

Penalty for words like “women”,
as in “women’s chess club”

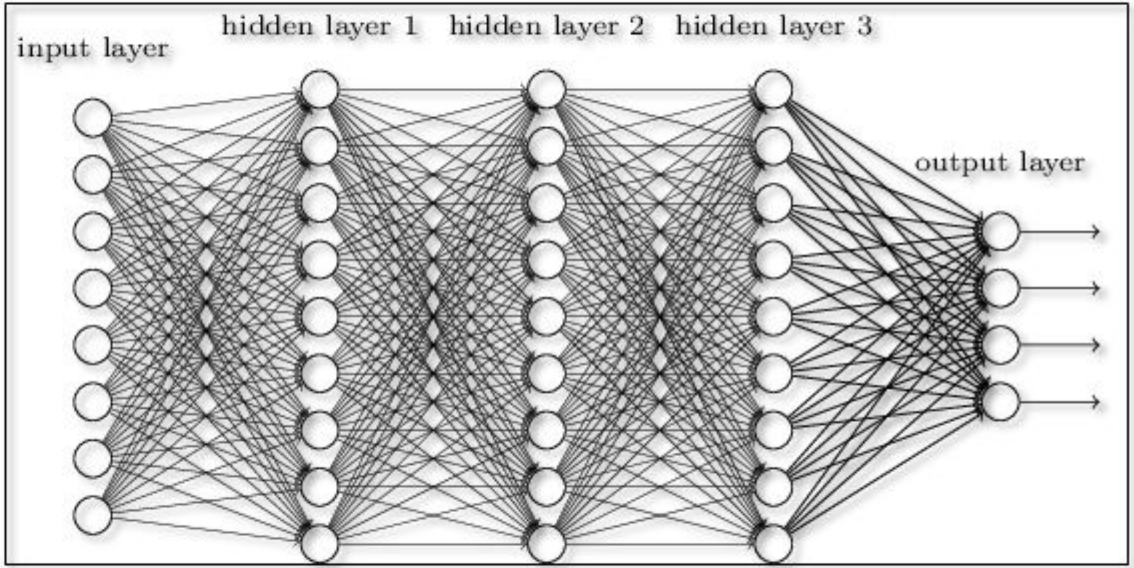
Eventually project was abandoned



Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

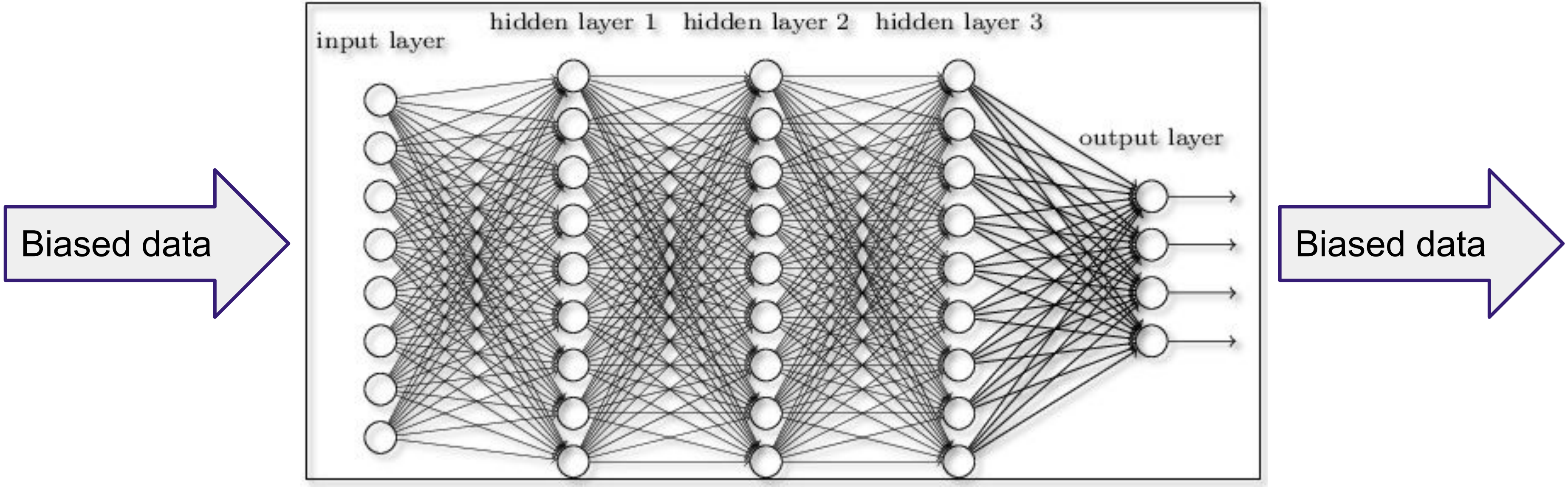
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Black magic



“Dog”

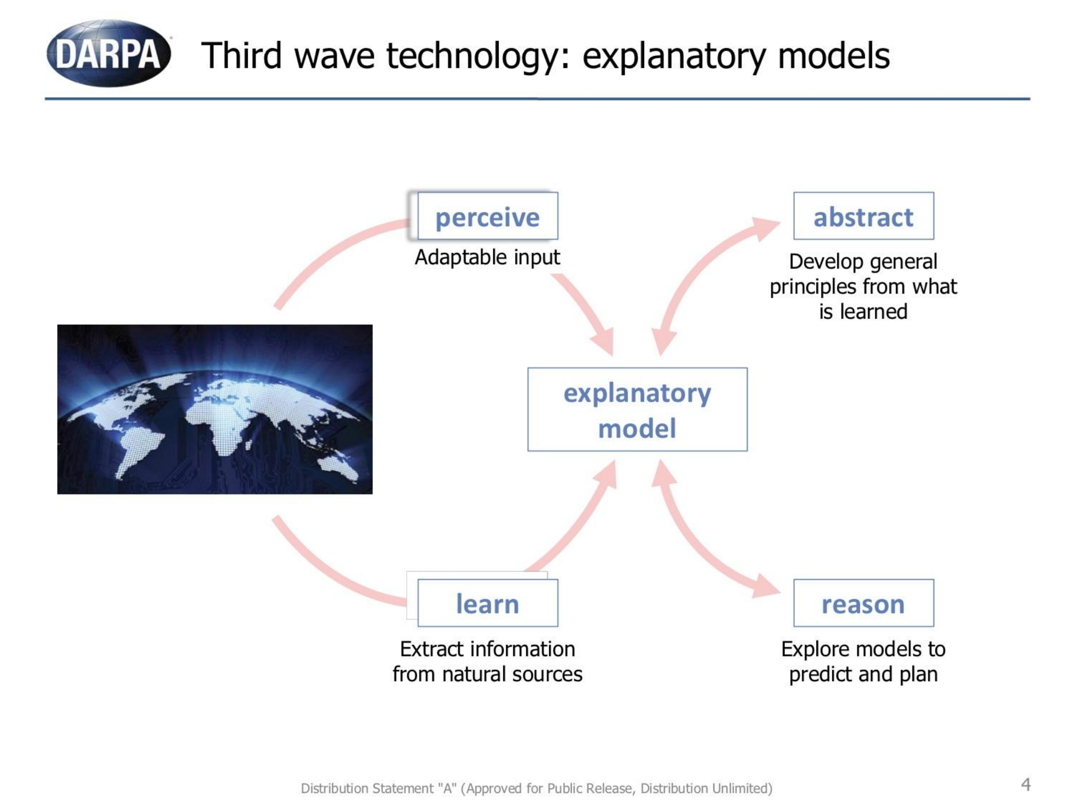
Black magic



Biased data

Biased data

What if we could ask, “Why?”



Third wave technology: explanatory models

perceive
Adaptable input

abstract
Develop general principles from what is learned

explanatory model

learn
Extract information from natural sources

reason
Explore models to predict and plan

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Who’s to blame?

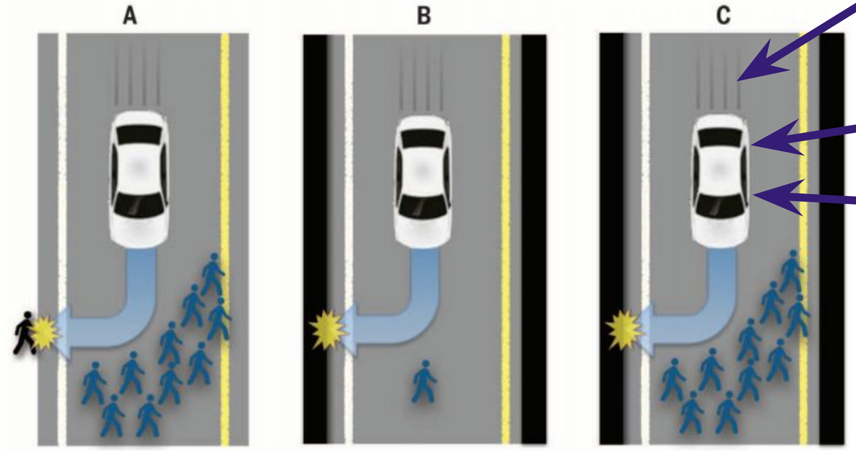


Fig. 1. Three traffic situations involving imminent unavoidable harm. The car must decide between (**A**) killing several pedestrians or one passerby, (**B**) killing one pedestrian or its own passenger, and (**C**) killing several pedestrians or its own passenger.

What if pavement was slipper?

What if brakes failed?

What if driver could have taken control?

Should a human driver in the same situation be blamed differently?

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576.

Different moral norms

Repairman vs. Repair robot

4 miners in a train that’s lost its breaks

Let the train run into a massive wall?

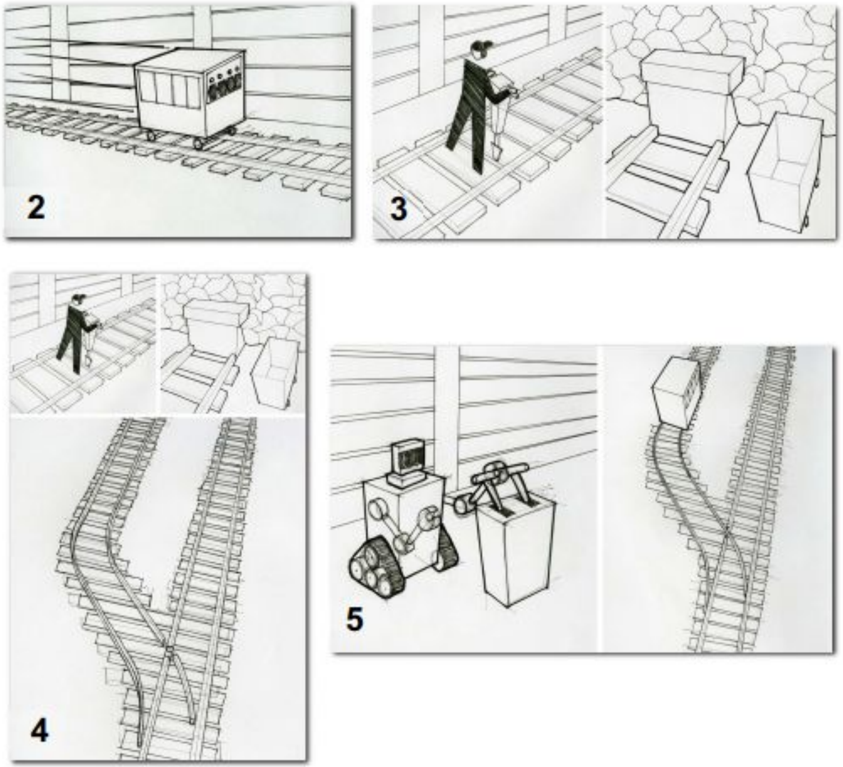
Divert train, killing a single worker?

Is diverting the train permissible?

How much blame to agent flipping switch?

Is the action morally wrong?

Fig. 3. Pictures 2 to 5 in the five-picture array condition. Picture 1 displayed the appropriate agent from Figure 1, and picture 5 showed this same agent again. All drawings ©Justin Finkenaur.



Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124). ACM.

Different moral norms

Repairman vs. Repair robot

4 miners in a train that’s lost its breaks

Let the train run into a massive wall?

Divert train, killing a single worker?

Is diverting the train permissible?

How much blame to agent flipping switch?

Is the action morally wrong?

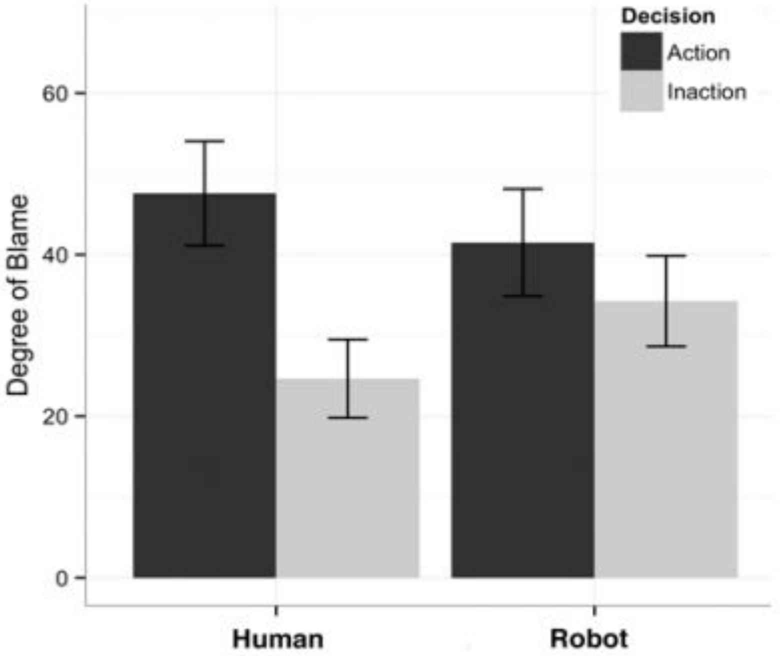


Figure 1. Rates of blame in Experiment 1 as a function of agent type and the agent’s decision (to divert the train or not)

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124). ACM.

What I didn’t do anything?

Can you blame the robot for doing nothing?

What if the robot just pleaded with you?

Can robots manipulate humans?

What if they can?

Robot protest

Robot builds a tower

Person tells robot to knock over tower

Robot cries

Tell the robot again?

- 4 of 10 do not re-issue command

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, *6*(3), 343-355.

Robot says, “No!”

https://qz.com/559432/robots-are-learning-to-say-no-to-human-orders-and-your-life-may-depend-on-it/

Briggs, G., & Scheutz, M. (2015, September). Sorry, I can’t do that”: Developing mechanisms to appropriately reject directives in human-robot interactions. In *2015 AAAI Fall Symposium Series*.

Midterm

See Melissa Duong (Mudd 3507) to pick up your exam

Stats

- Average: 50.176 / 60
- SD: 5.68

Up next

Uncertainty

Chapters 13 & 14