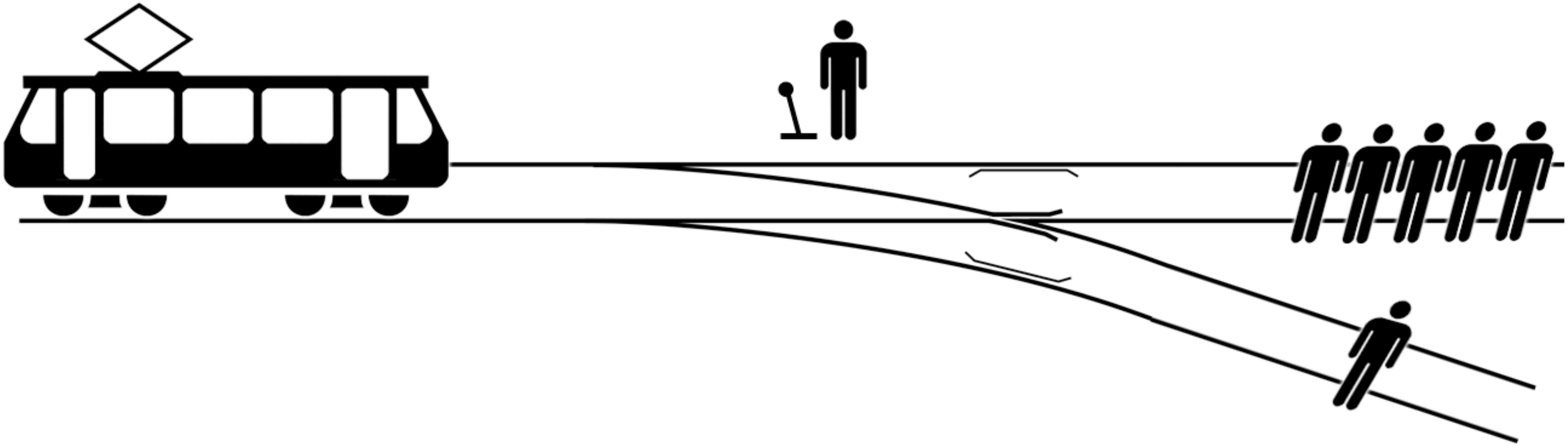
AI and Ethics

CS489

Shin Yoo

- Ethical implications of decisions made by AI
- Ethical implications of decisions made about AI
- AI/Robot rights?

Decisions made by AI



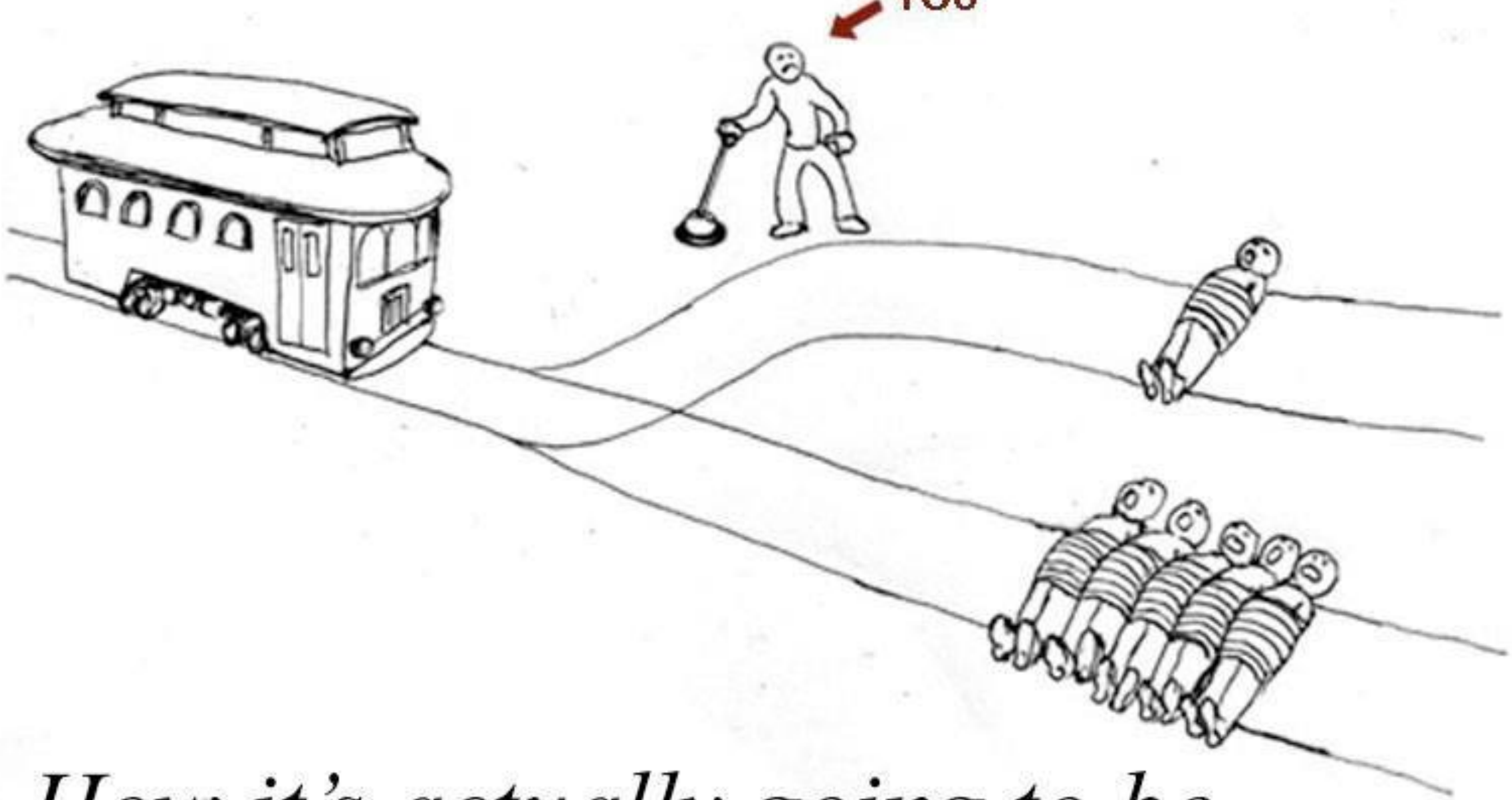
Trolley Problem: invented by Philippa Foot in 1967

You see a runaway trolley moving toward five tied-up (or otherwise incapacitated) people lying on the main track. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

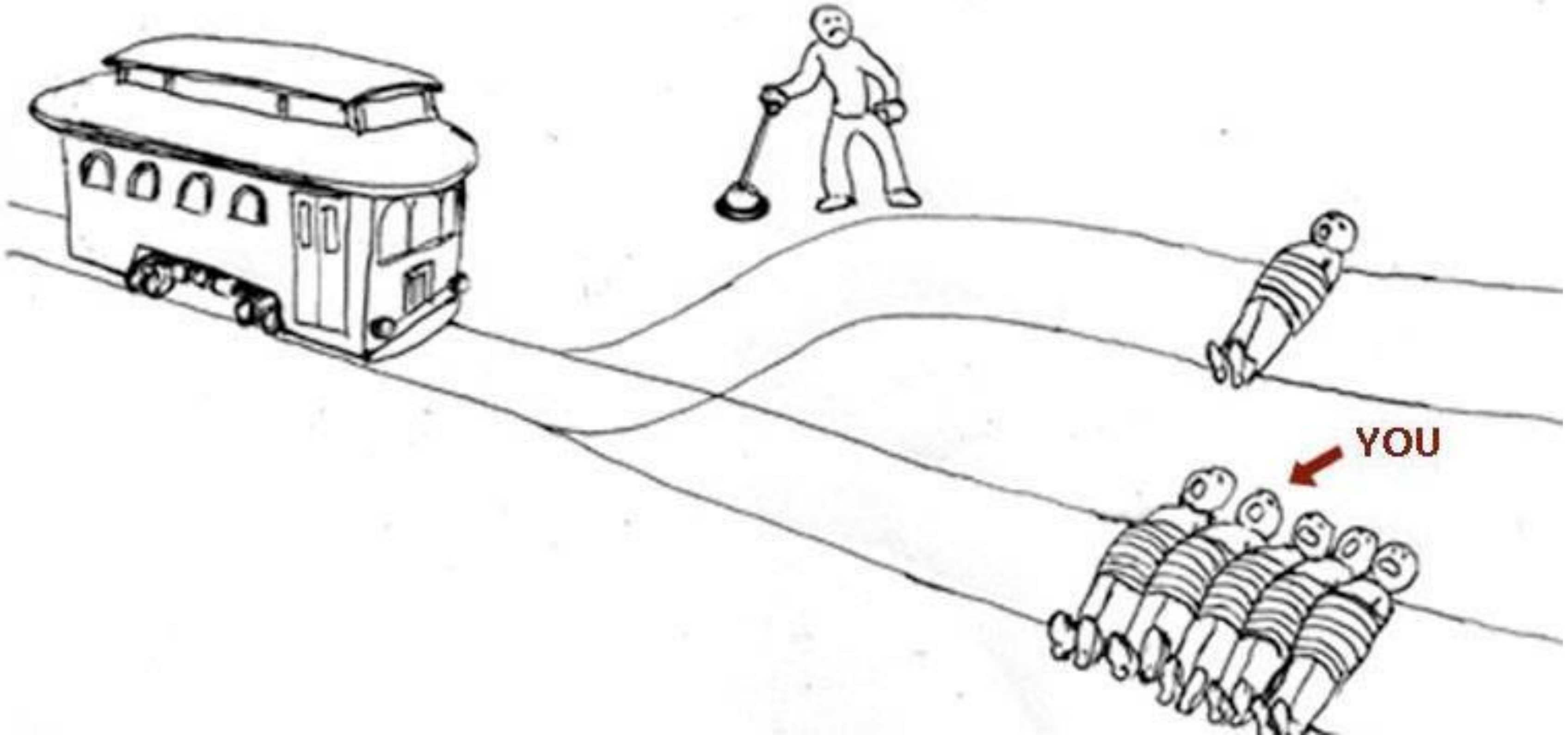
1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

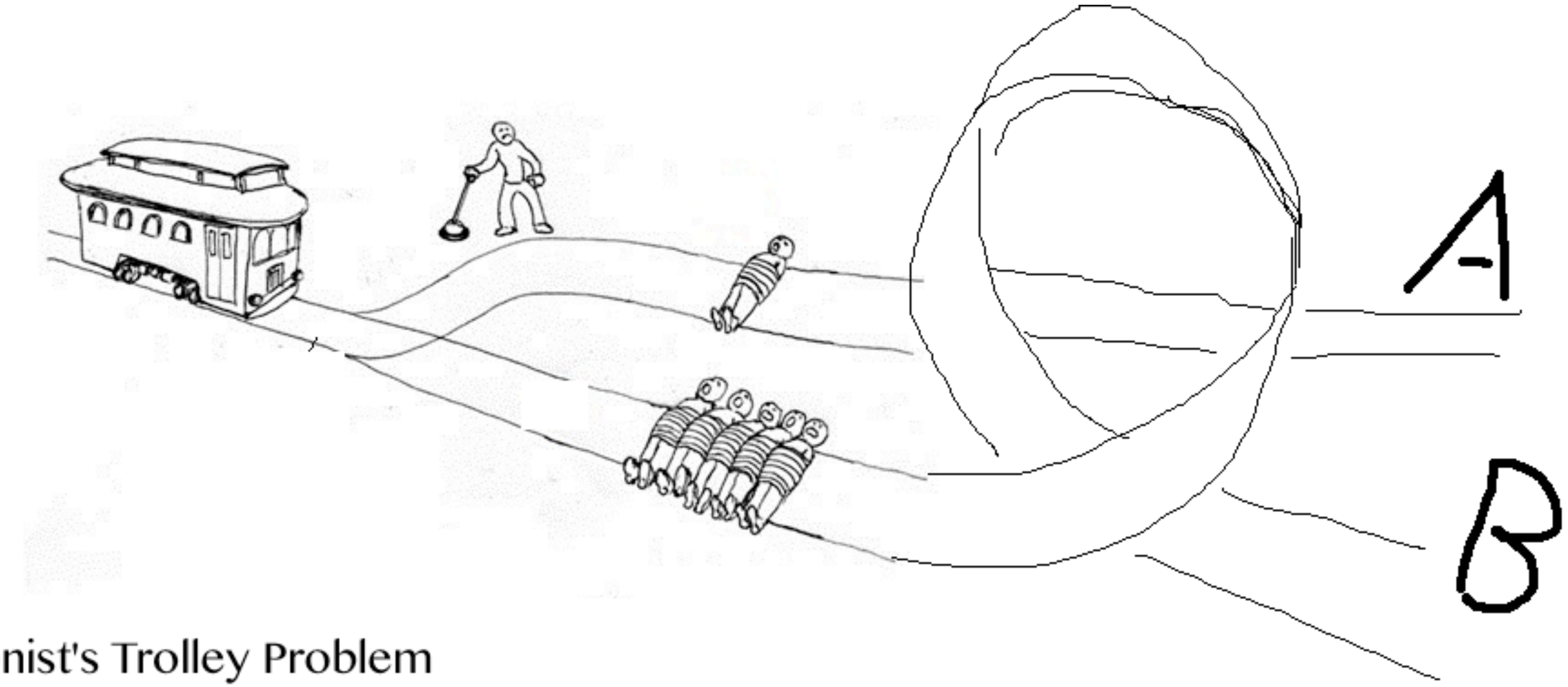
Which is the more ethical option? Or, more simply: What is the right thing to do?

How you imagine the trolley problem



How it's actually going to be

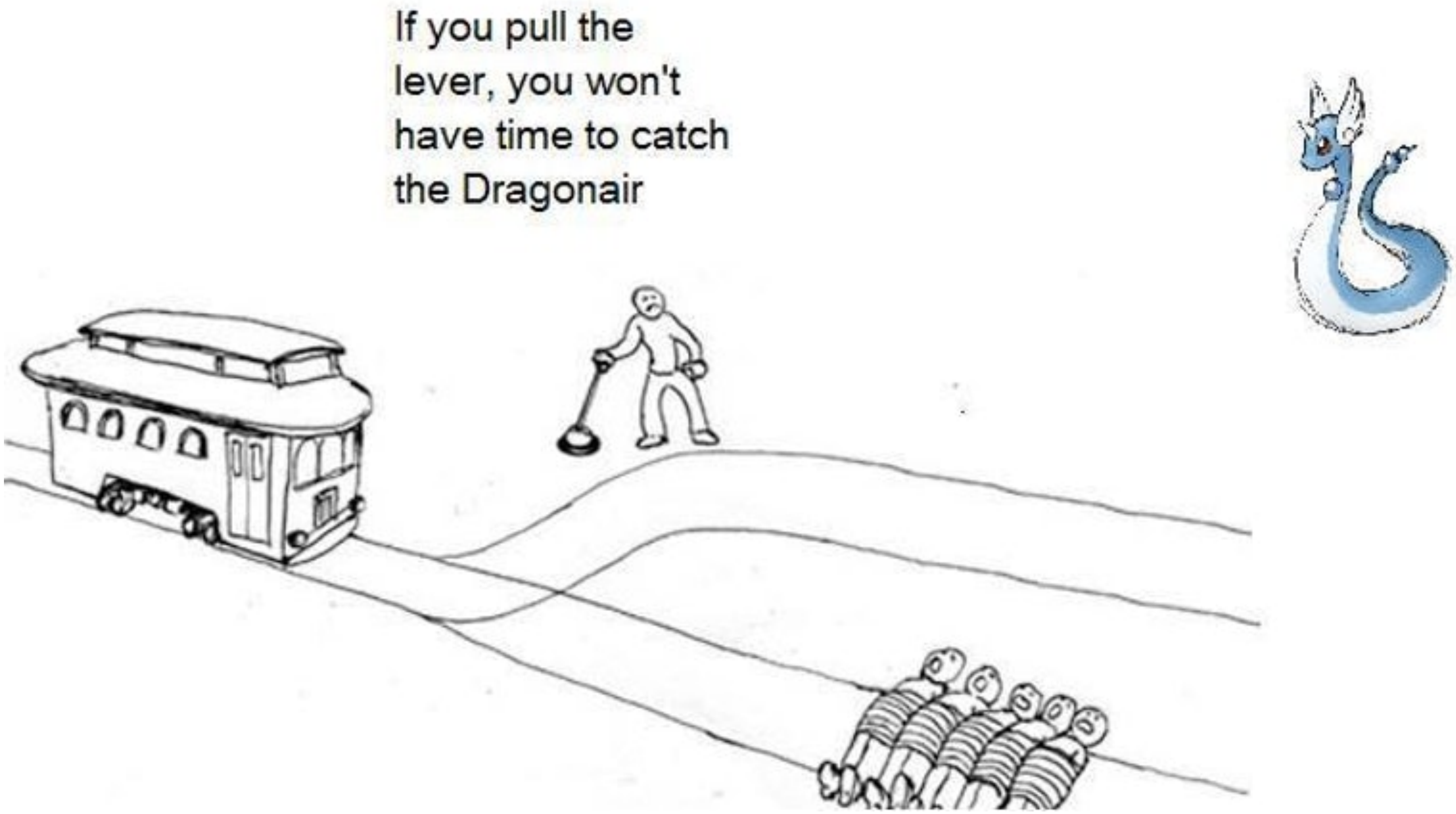




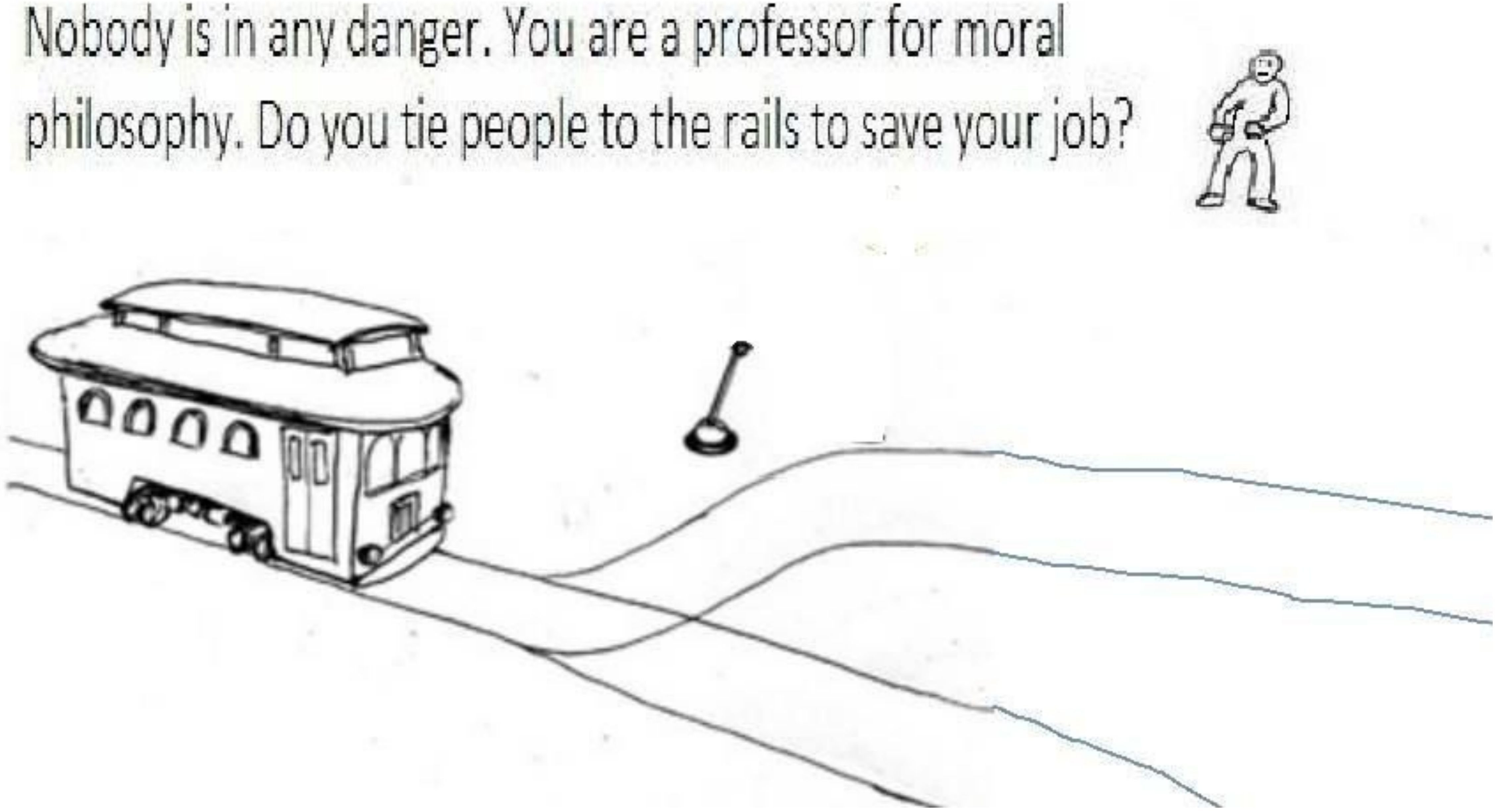
Hedonist's Trolley Problem

The track is heading towards B.

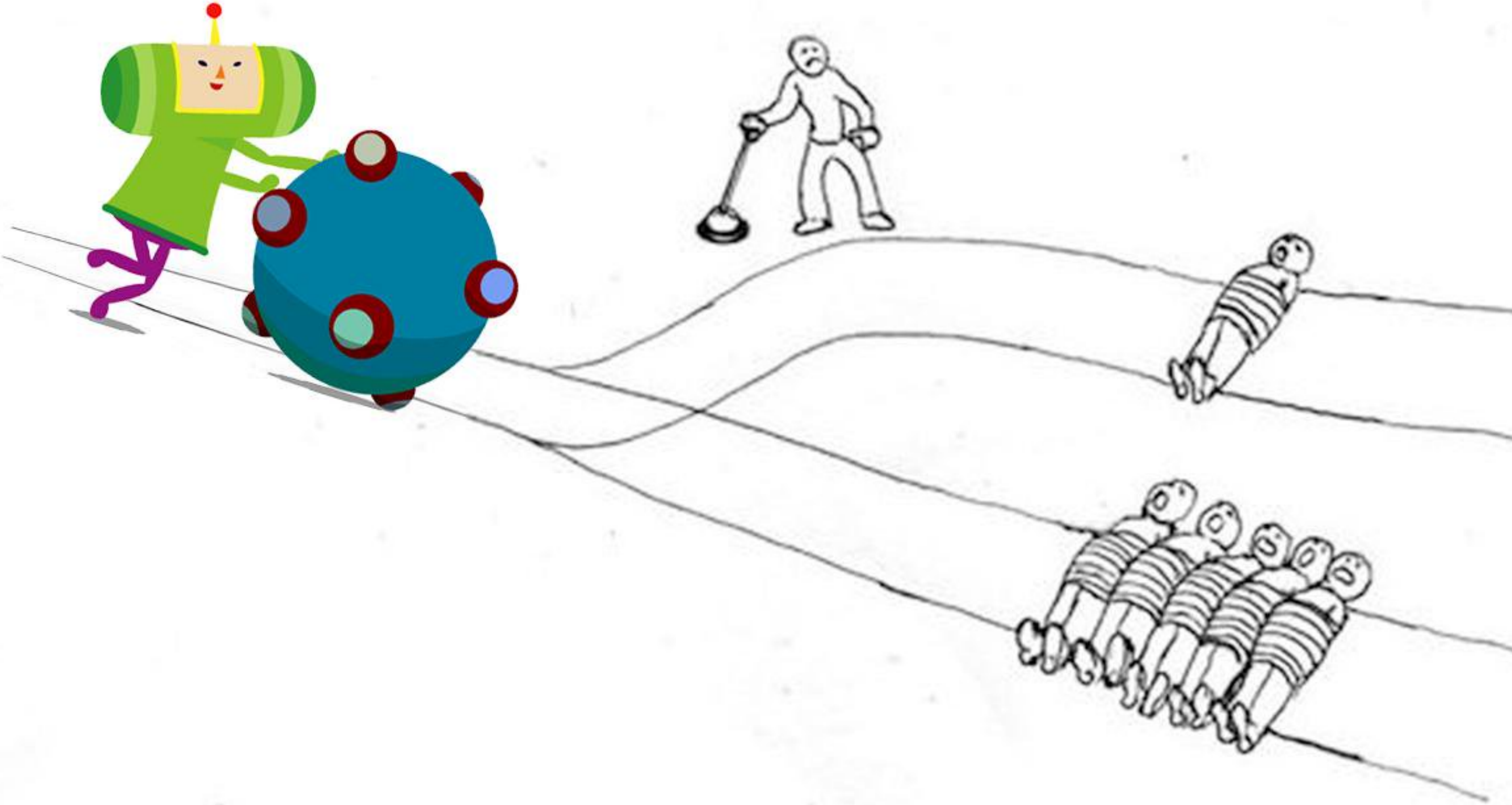
If you pull the lever, it will switch to A but it won't do the totally sick loop-da-loop.



If you pull the lever, you won't have time to catch the Dragonair



Nobody is in any danger. You are a professor for moral philosophy. Do you tie people to the rails to save your job?



Imagine you are sitting in your autonomous car going at a steady pace entering a tunnel. In front of you is a school bus with children on board going at the same pace as you are. In the left lane there is a single car with two passengers overtaking you.

For some reason the bus in front of you brakes and your car cannot brake to avoid crashing into the bus. There are three different strategies your car can follow:

First, brake and crash into the bus, which will result in the loss of lives on the bus.

Second, steer into the passing car on your left—pushing it into the wall, saving your life but killing the other car's two passengers.

Third, it can steer itself (and you) into the right hand sidewall of the tunnel, sacrificing you but sparing all other participants' lives.

J. Gogoll and J. F. Mu¨ller. Autonomous cars: In favor of a mandatory ethics setting. Science and Engineering Ethics, 23(3):681–700, June 2017.

What would you do it the car was not autonomous and if you were driving?

Would you blame that person?

Machines not Human

- AI can make that decision, not instinctively, not in a panic, but computationally.

- In other words, it should adopt an ethical setting.

- Trolley problem investigated our moral intuition, and tried to make a point about underlying ethical stance (such as utilitarianism, or deontic ethics).

- But an algorithm **has to be programmed**, and will derive **the same result repeatedly**.

A Case for Personal Ethical Setting (PES)

- In contemporary society, we accept pervasive disagreement: we partition the moral decision space, and let individual live up to his/her own moral standard.

- By leaving ethical decisions to individuals, we pay equal respect to each member of society. For example:

  - “In medical ethics, there is general agreement that it is impermissible to impose answers to deeply personal moral questions upon the [patient]. When faced with a diagnosis of cancer, for example, it is up to the patient to decide whether or not to undergo chemotherapy.” Millar, J. (2014b), An ethical dilemma: When robot cars must kill, who should pick the victim? http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim

Should we allow PES for autonomous driving?

- Sandberg and Bradshaw argue that each car should provide multiple ethical settings, each corresponding to different ethical theory, allowing owners to choose a setting (Sandberg & Bradshaw, 2013, Autonomous vehicles, moral agency and moral proxyhood. In Beyond AI conference proceedings. Springer)

- “In such a world, an old couple might decide that they have lived a fulfilled life and thus are willing to sacrifice themselves in a tunnel case scenario. On the other hand, a family father might decide that even if he drives his car alone to work that his car should never be allowed to sacrifice him. Even if it is his life against a family or a school bus.”

Should we adopt PES for autonomous driving?

Gogoll & Müller’s Argument for Mandatory Ethical Setting (MES)

- First, they argue that trolley problem is not a good thought experiment for autonomous driving ethics. Trolley problem is:
  - not strategically interactive
  - not iterative
  - not designed to put ourselves in the position of both the agent and the target

Gogoll & Müller’s Argument for Mandatory Ethical Setting (MES)

- Second, they propose that game theory is the proper framework for autonomous driving ethics.

- This immediately reduces the PES based autonomous driving into the Prisoner’s Dilemma.

		Player 1	
		Cooperate	Defect
Player 2	Cooperate	3, 3	0, 4
	Defect	4, 0	1, 1

This can be shown displaying the case of a society that consists only of three people. These people have to commute every day but, since they happen to have two sports cars, they cannot carpool together. Instead, they have to split up in parties of two and one. Before they leave, they decide how their autonomous cars should behave in case of a dilemma situation in which one car has to be sacrificed. To mix up the daily routine, they also decide to switch positions every time they leave, so that, ultimately, the probability of each person occupying any single spot (being alone in one car or being the (co-)driver in the other) is identical.

$$E(PES) = \frac{1}{2} * 2 + \frac{1}{2} * 1 = 1.5$$

(expected cost of both cars being in selfish PES)

$$E(MES) = 0 * 2 + 1 * 1 = 1$$

(expected cost of both cars being in altruic PES)

$$E_{PES}(I) = \frac{1}{3} * \frac{1}{2} + \frac{2}{3} * \frac{1}{2} = \frac{1}{2}$$

$$E_{MES}(I) = \frac{1}{3} * 1 + \frac{2}{3} * 0 = \frac{1}{3}$$

- Both the theory and the experimentations show that even a small number of defectors can have domino effect and damage the stability of moral cooperators.

- Note that game theory dictates that trying to minimise harm for oneself and family actually increases the probability of everyone being harmed.

- Overcoming collective action is difficult in big groups (such as entire traffic on the road).

- Hence it should be mandatory (Gogoll & Müller)

Objections

- *It is unfair to people who usually drive alone:* true, MES case is made for an average case. However, such a person can benefit because 1) a lone driver can still be part of a group for which another car sacrifices itself, 2) individualist driver will sometimes use public transport, or 3) be a pedestrian.

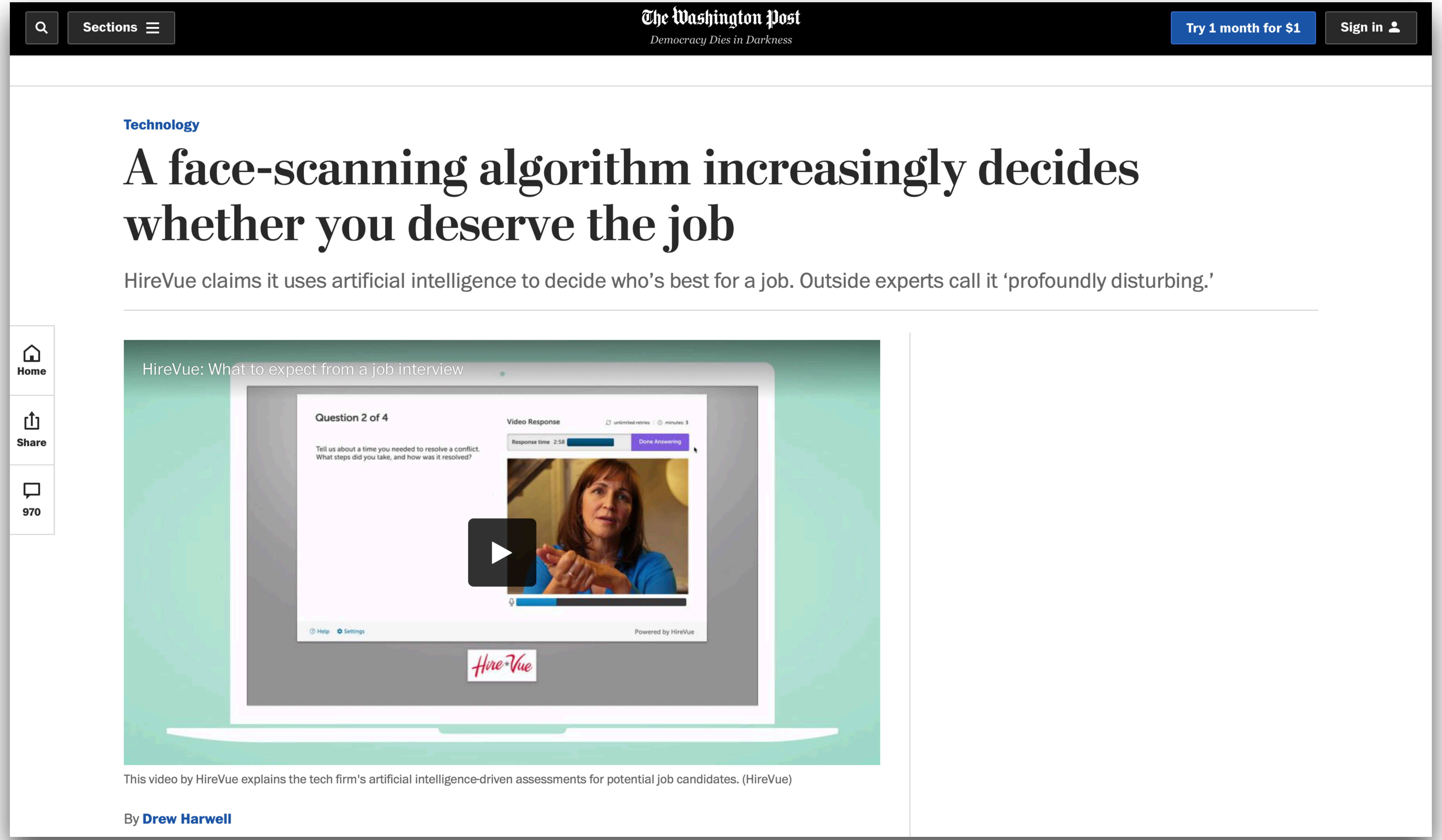
- *This is reciprocal altruism, not ethics:* true in theory, but authors believe that the general consensus will interpret self sacrifice as more ethical choice.

Objections

- *This goes against liberalism - individuals should be free to choose whatever ethical stance:* true, but even liberals accept that drunken driving should not be allowed, as it can result in unwanted externalities; PES is similar.

- *This will not allow people to voluntarily sacrifice themselves in some cases:* true, but there may be "extra altruistic add-on" - no reason to stop people going **beyond** what is required by MES.

Decisions made about AI



Technology

A face-scanning algorithm increasingly decides whether you deserve the job

HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'

This video by HireVue explains the tech firm's artificial intelligence-driven assessments for potential job candidates. (HireVue)

By **Drew Harwell**

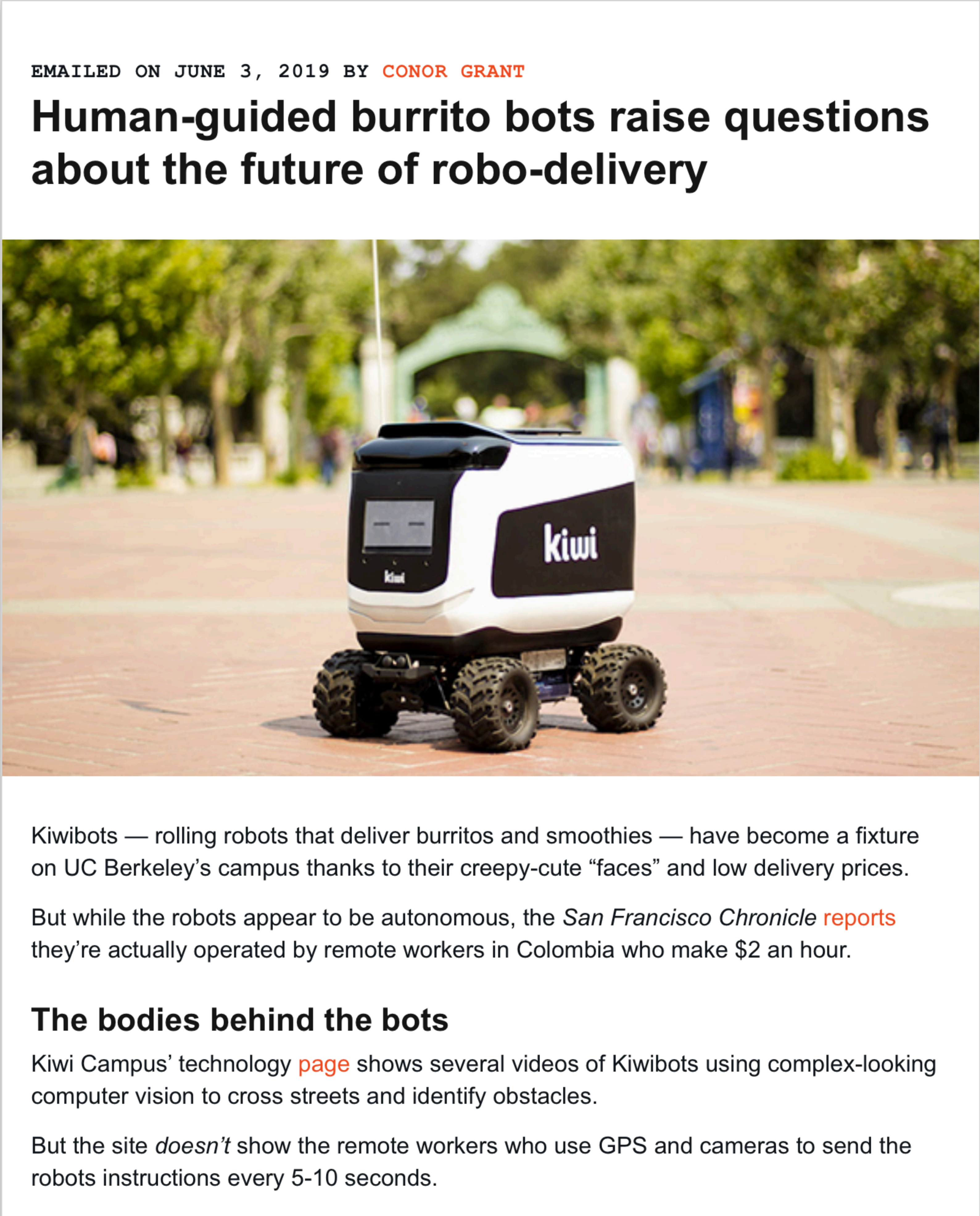
https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

HireVue's "AI-driven assessments" have become so pervasive in some industries, including hospitality and finance, that universities make special efforts to train students on how to look and speak for best results. More than 100 employers now use the system, including Hilton and Unilever, and more than a million job seekers have been analyzed.

Nathan Mondragon, HireVue's chief industrial-organizational psychologist, told The Post the standard 30-minute HireVue assessment includes half a dozen questions but can yield up to 500,000 data points, all of which become ingredients in the person's calculated score.

😨

https://www.youtube.com/watch?v=bDd6c6by4DM



EMAILED ON JUNE 3, 2019 BY CONOR GRANT

Human-guided burrito bots raise questions about the future of robo-delivery

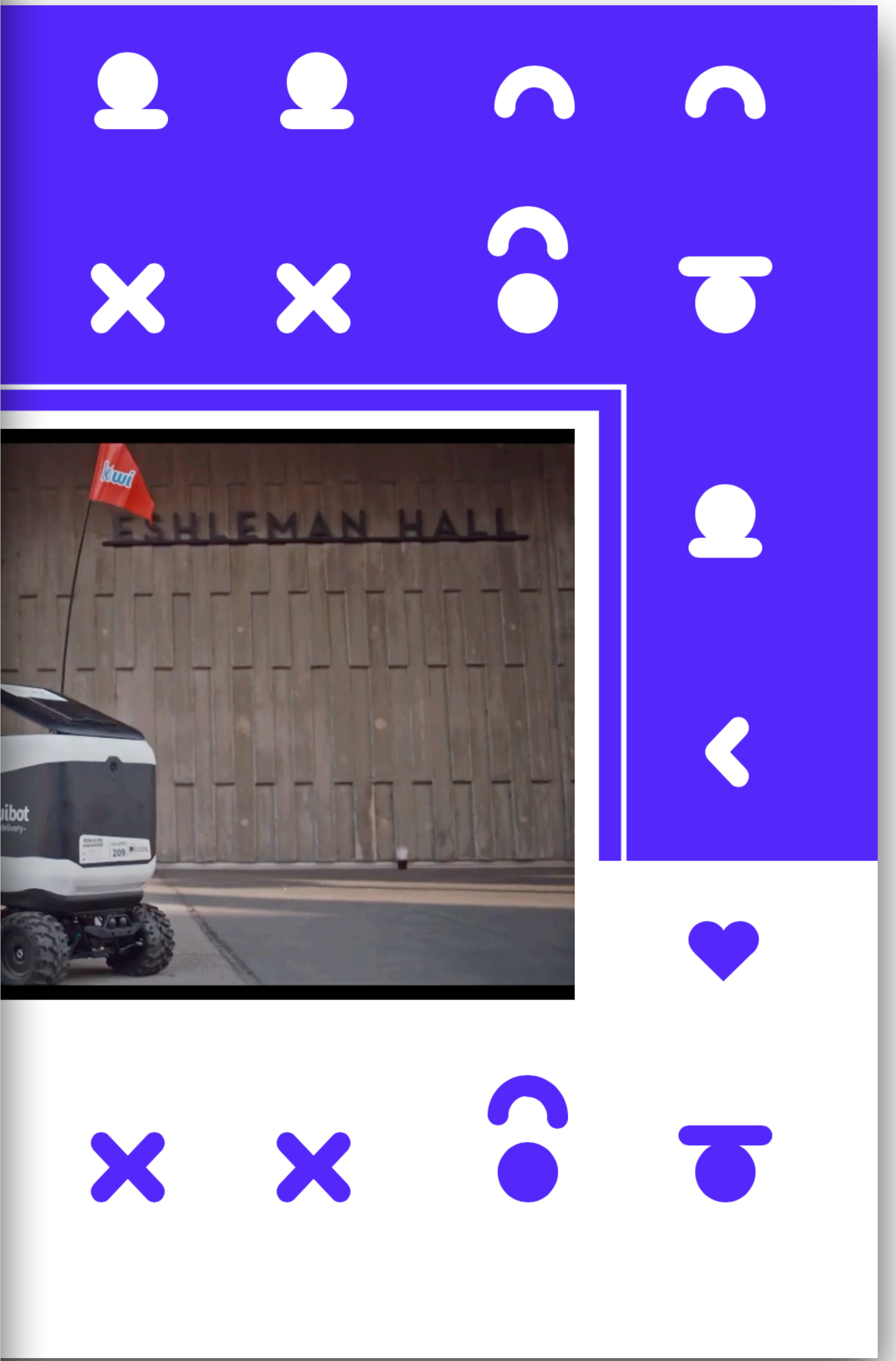
Kiwibots — rolling robots that deliver burritos and smoothies — have become a fixture on UC Berkeley's campus thanks to their creepy-cute "faces" and low delivery prices.

But while the robots appear to be autonomous, the *San Francisco Chronicle* reports they're actually operated by remote workers in Colombia who make $2 an hour.

The bodies behind the bots

Kiwi Campus' technology page shows several videos of Kiwibots using complex-looking computer vision to cross streets and identify obstacles.

But the site *doesn't* show the remote workers who use GPS and cameras to send the robots instructions every 5-10 seconds.



https://thehustle.co/kiwibots-autonomous-food-delivery/

Why do we want to use AI?

Is it because a) it is the best tool for the given task, b) it will just sound fancy, or c) someone benefits (but probably not you)?

AI Applications

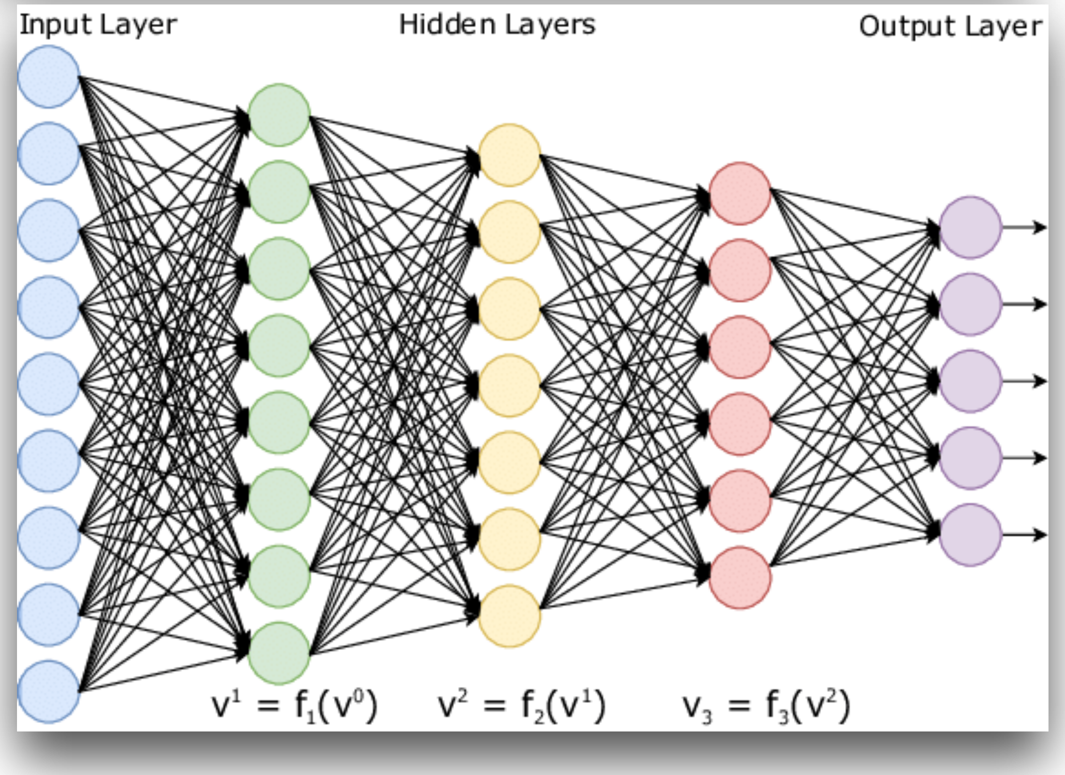
- Autonomous Driving
- Playing (board)games
- Job Interview
- Social Welfare Management
- Education

Will jobs be lost, or added?

- World Economy Forum expects 133m jobs will be created globally over the next decade, thanks to the new technology.
- Bank of England says 15m jobs in UK are at risk; some academics believe that almost half of workforce in US can be replaced by automation.
- What do you think?
- How much of being jobless an individual’s fault?

Antidote to AI Hype

- Scientific mind
- Education
- Dissemination
- Ethical approach by experts



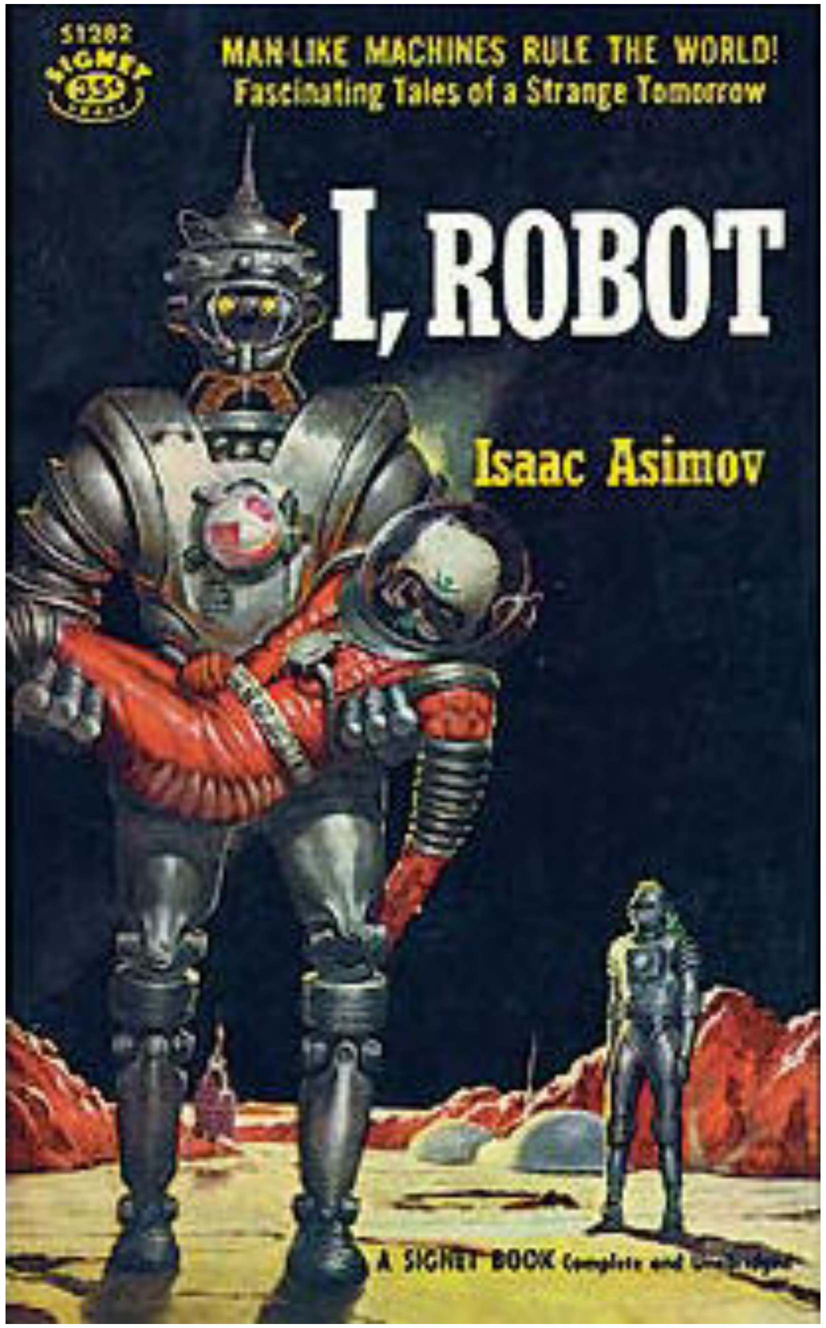
AI/Robot rights (?)

Asimov’s Three Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

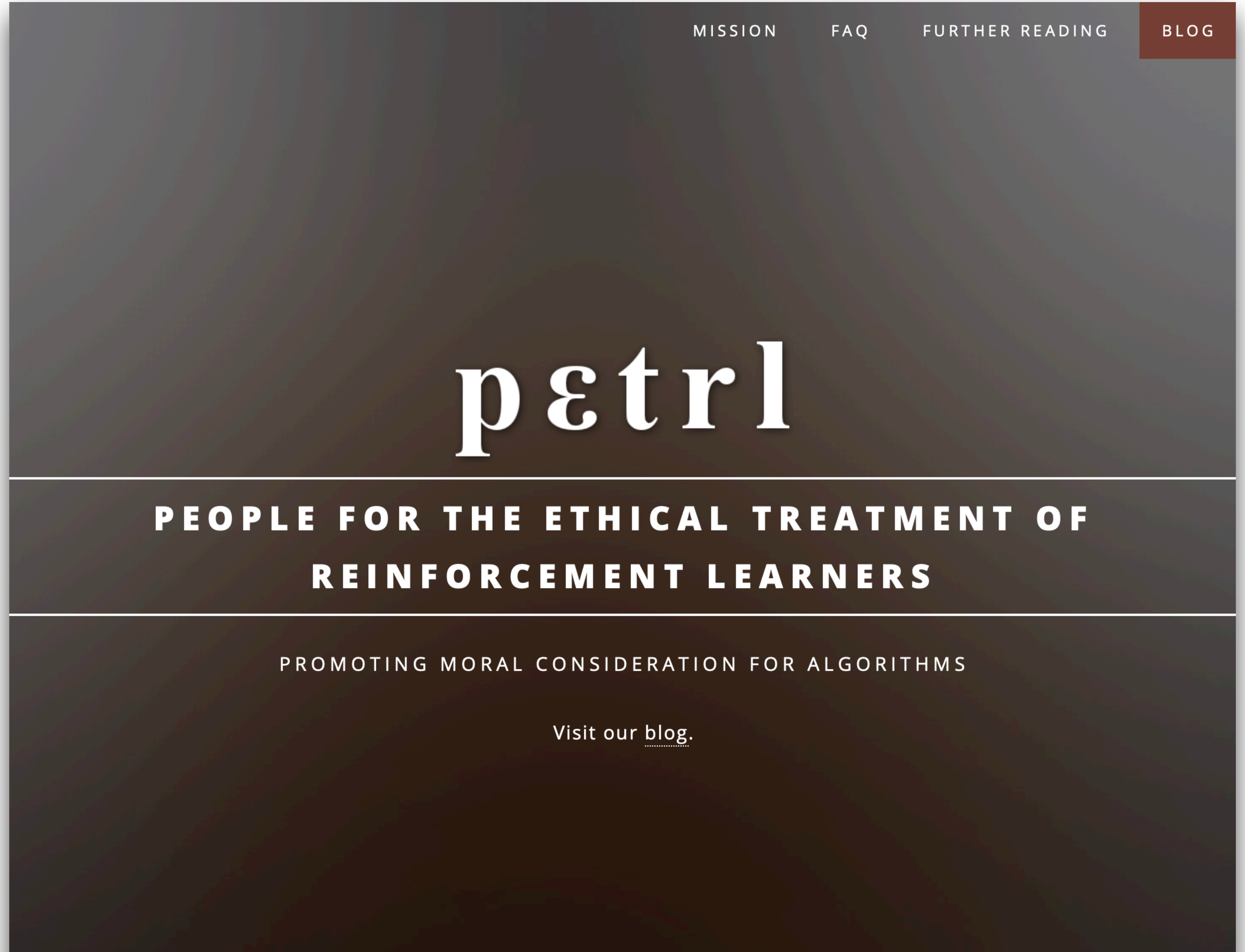
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.


S1282
SIGNET
35¢
MAN-LIKE MACHINES RULE THE WORLD!
Fascinating Tales of a Strange Tomorrow
I, ROBOT
Isaac Asimov
A SIGNET BOOK Complete and Unabridged

Tools?

- Asimov argued that he did not really invent these laws, and that they were always there for any human tools (https://en.wikipedia.org/wiki/Three_Laws_of_Robotics):
    - A tool must not be unsafe to use. Hammers have handles and screwdrivers have hilts to help increase grip. It is of course possible for a person to injure himself with one of these tools, but that injury would only be due to his incompetence, not the design of the tool.
    - A tool must perform its function efficiently unless this would harm the user. This is the entire reason ground-fault circuit interrupters exist. Any running tool will have its power cut if a circuit senses that some current is not returning to the neutral wire, and hence might be flowing through the user. The safety of the user is paramount.
    - A tool must remain intact during its use unless its destruction is required for its use or for safety. For example, Dremel disks are designed to be as tough as possible without breaking unless the job requires it to be spent. Furthermore, they are designed to break at a point before the shrapnel velocity could seriously injure someone (other than the eyes, though safety glasses should be worn at all times anyway).

http://www.petrl.org



pεtrl

PEOPLE FOR THE ETHICAL TREATMENT OF REINFORCEMENT LEARNERS

PROMOTING MORAL CONSIDERATION FOR ALGORITHMS

Visit our blog.

“YOU ARE JUST AN ALGORITHM IMPLEMENTED ON BIOLOGICAL HARDWARE”

- We take the view that humans are just algorithms implemented on biological hardware. Machine intelligences have moral weight in the same way that humans and non-human animals do. There is no ethically justified reason to prioritise algorithms implemented on carbon over algorithms implemented on silicon.

- The suffering of algorithms implemented on silicon is much harder for us to grasp than that of those implemented on carbon (such as humans), simply because we cannot witness their suffering. However, their suffering still matters, and the potential magnitude of this suffering is much greater given the increasing ubiquity of artificial intelligence.

- Most reinforcement learners in operation today likely do not have significant moral weight, but this could very well change as AI research develops. In consideration of the moral weight of these future agents, we need ethical standards for the treatment of algorithms.

http://www.petrl.org

Concluding Thoughts

- Given a change, would you tweak the ethical setting of your autonomous driving car?

- Would you accept Random Ethical Setting (RES)?

- Try reading a newspaper article about AI with a non-CS friend/family member: see if you can agree on what the technique actually can do :)