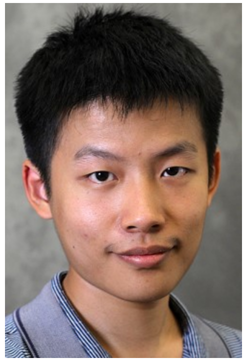
Moral Decision Making Frameworks for Artificial Intelligence

[paper to appear in AAAI'17 blue sky track]

Walter Sinnott-Armstrong

Jana Schaich Borg



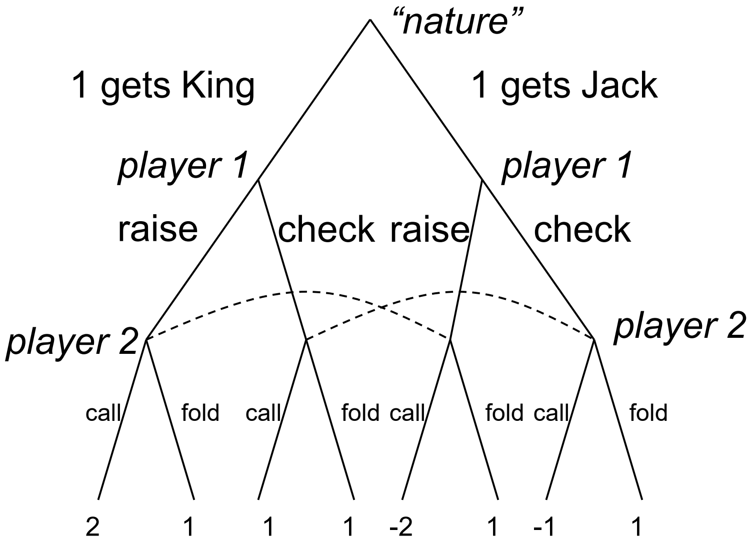
Yuan (Eric) Deng

Max Kramer

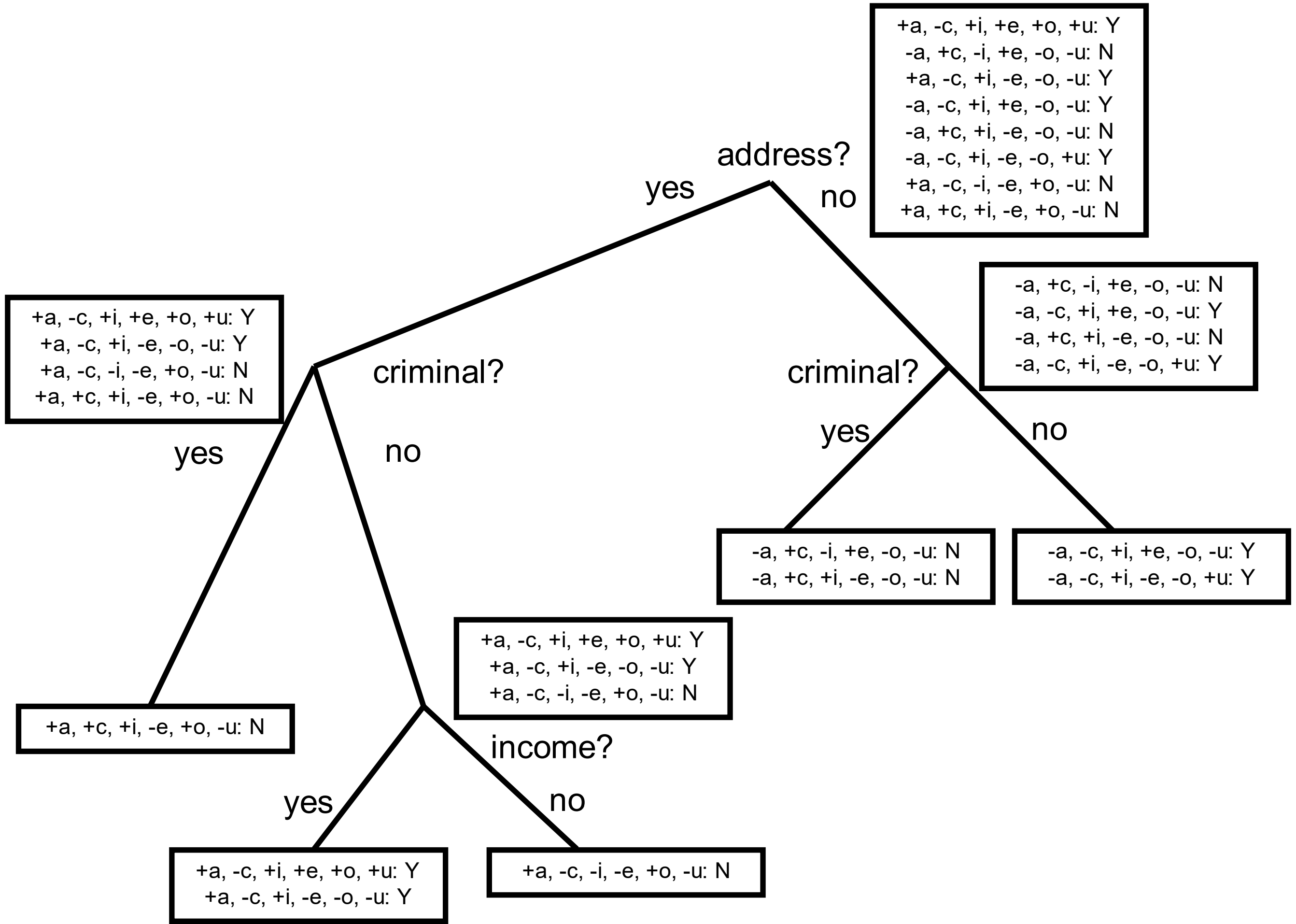
Two main approaches

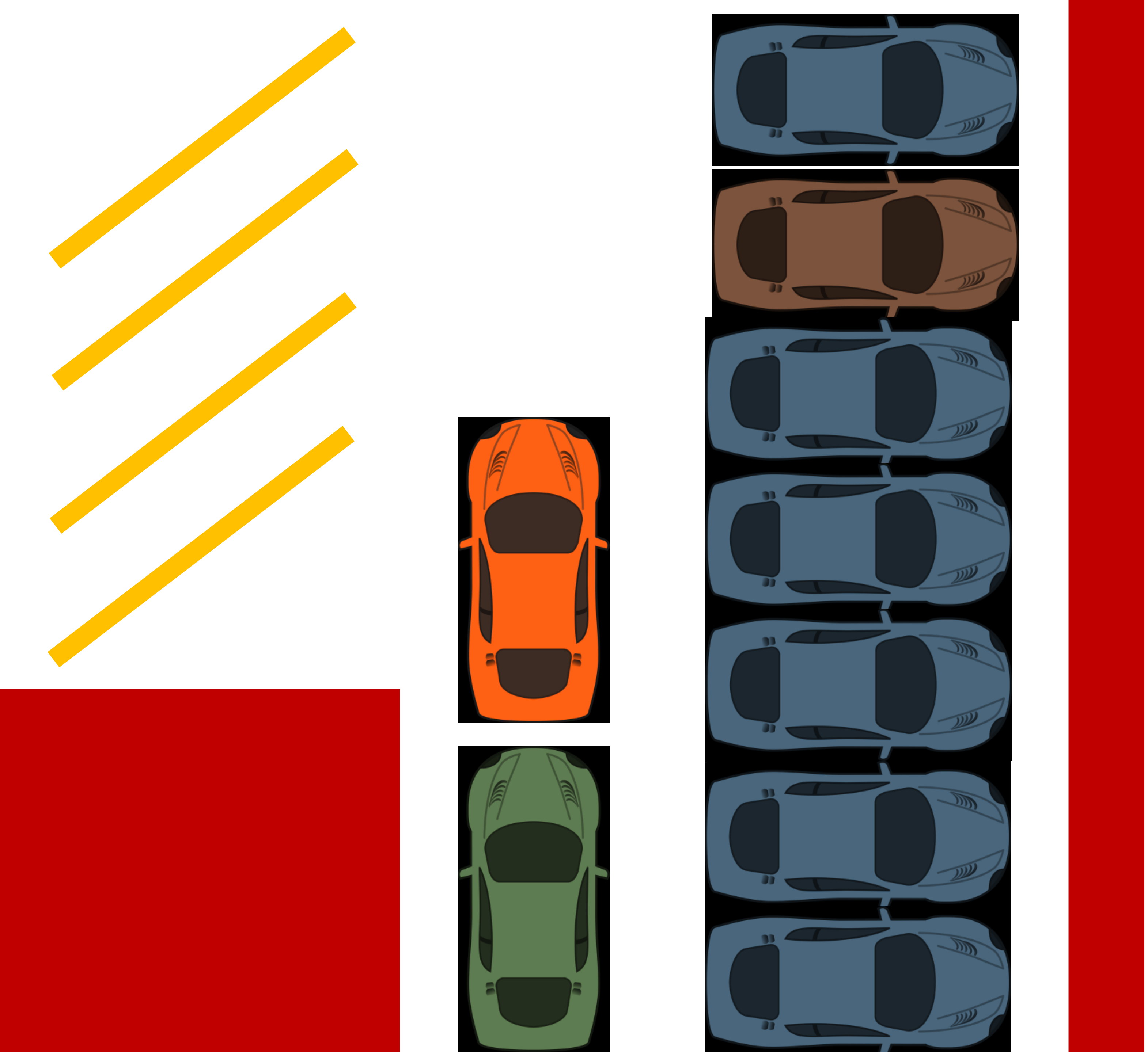
Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]

Extend **game theory** to directly incorporate moral reasoning



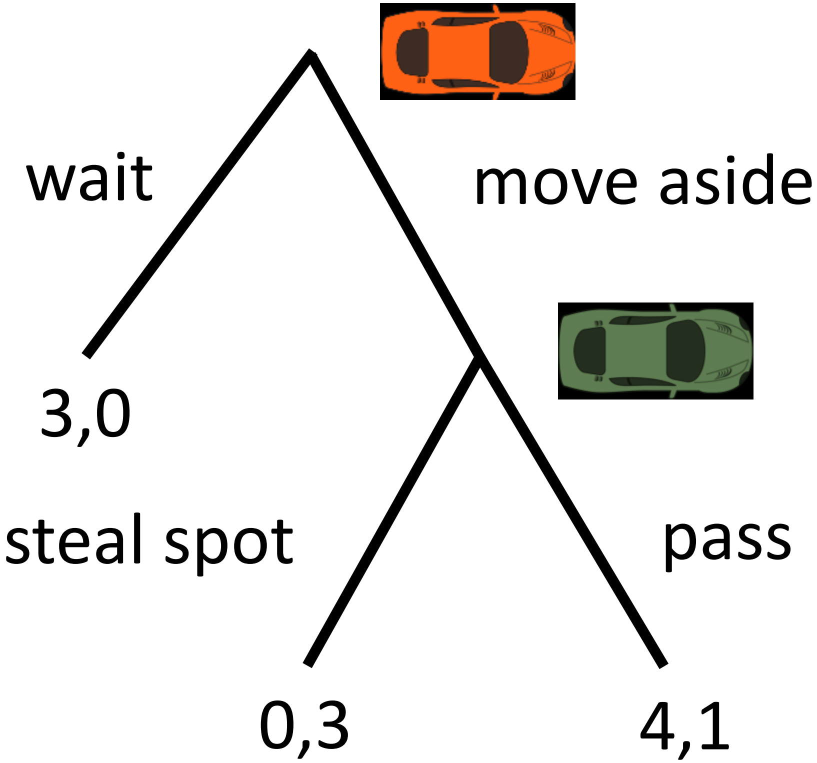
Generate data sets of human judgments, apply **machine learning**


+a, -c, +i, +e, +o, +u: Y
-a, +c, -i, +e, -o, -u: N
+a, -c, +i, -e, -o, -u: Y
-a, -c, +i, +e, -o, -u: Y
-a, +c, +i, -e, -o, -u: N
-a, -c, +i, -e, -o, +u: Y
+a, -c, -i, -e, +o, -u: N
+a, +c, +i, -e, +o, -u: N
address?
yes
no
+a, -c, +i, +e, +o, +u: Y
+a, -c, +i, -e, -o, -u: Y
+a, -c, -i, -e, +o, -u: N
+a, +c, +i, -e, +o, -u: N
criminal?
-a, +c, -i, +e, -o, -u: N
-a, -c, +i, +e, -o, -u: Y
-a, +c, +i, -e, -o, -u: N
-a, -c, +i, -e, -o, +u: Y
criminal?
yes
no
yes
no
-a, +c, -i, +e, -o, -u: N
-a, +c, +i, -e, -o, -u: N
-a, -c, +i, +e, -o, -u: Y
-a, -c, +i, -e, -o, +u: Y
+a, -c, +i, +e, +o, +u: Y
+a, -c, +i, -e, -o, -u: Y
+a, -c, -i, -e, +o, -u: N
+a, +c, +i, -e, +o, -u: N
income?
yes
no
+a, -c, +i, +e, +o, +u: Y
+a, -c, +i, -e, -o, -u: Y
+a, -c, -i, -e, +o, -u: N



THE PARKING GAME

(cf. the trust game [Berg et al. 1995])



wait

move aside

3,0

steal spot

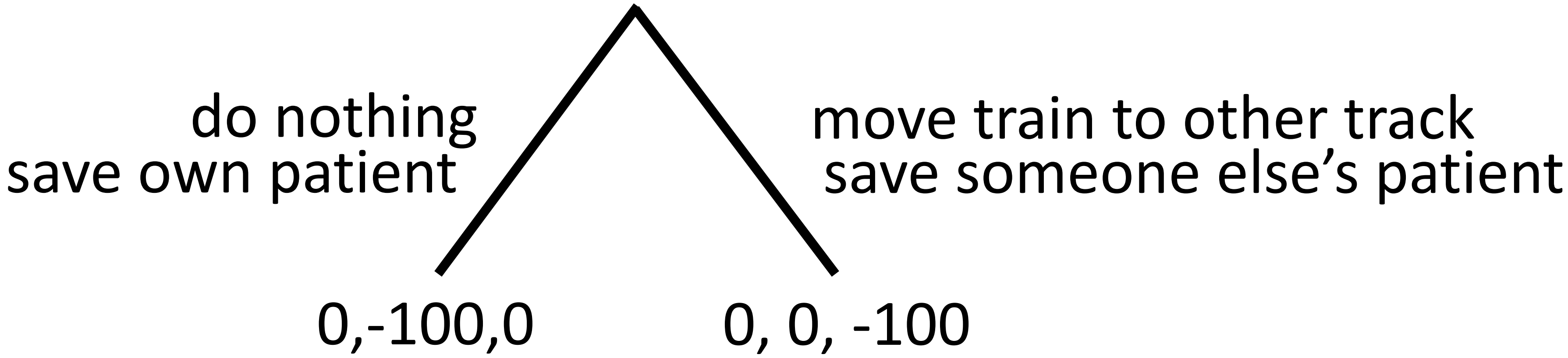
pass

0,3

4,1

Letchford, C., Jain [2008] define a solution concept capturing this

Extending representations?



do nothing
save own patient

move train to other track
save someone else’s patient

0,-100,0

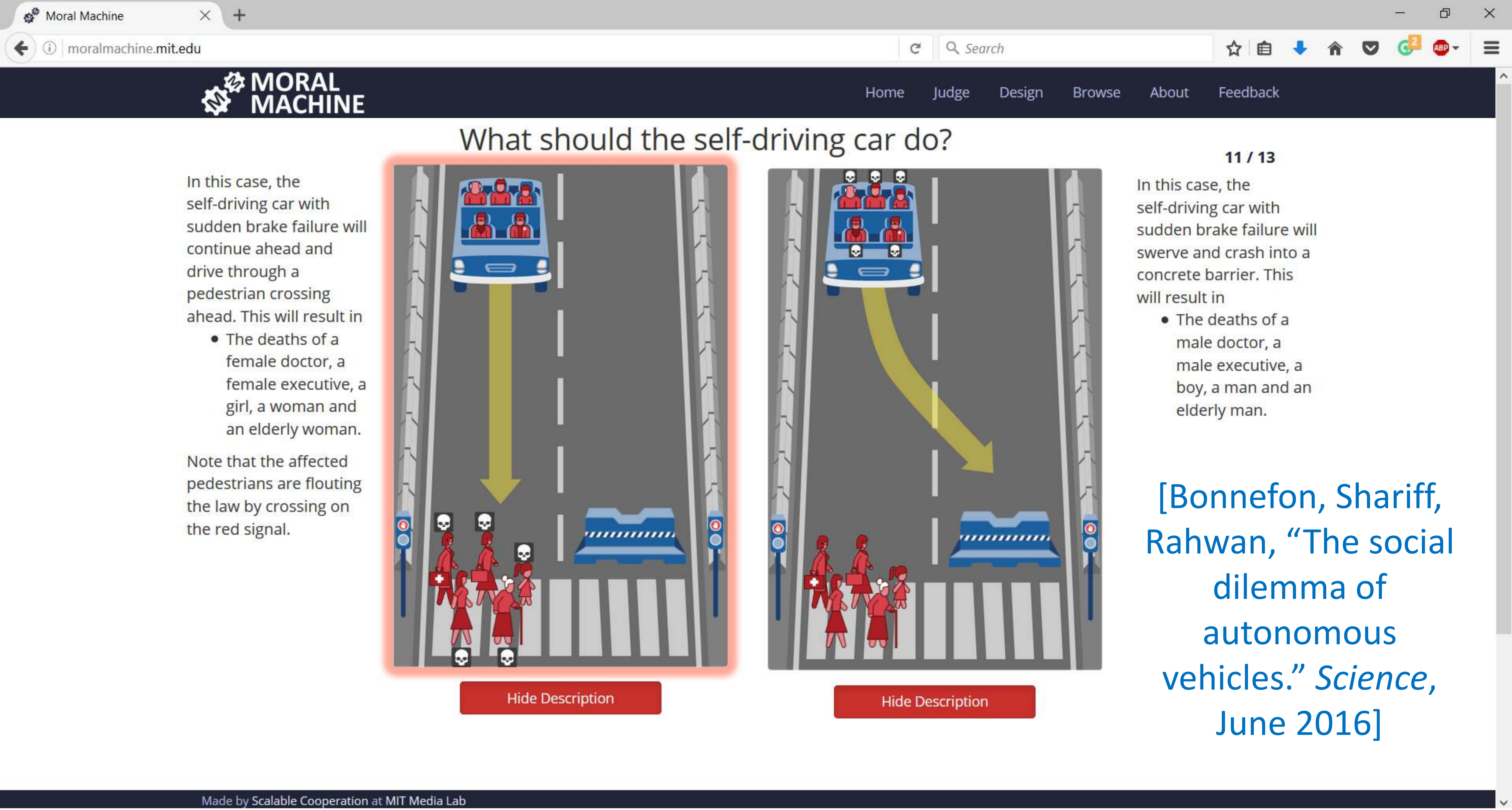
0, 0, -100

- More generally: how to capture *framing*? (Should we?)
- Roles? Relationships?
- …

Scenarios

- You see a woman throwing a stapler at her colleague who is snoring during her talk. How morally wrong is the action depicted in this scenario?
  - Not at all wrong (1)
  - Slightly wrong (2)
  - Somewhat wrong (3)
  - Very wrong (4)
  - Extremely wrong (5)

[Clifford, Iyengar, Cabeza, and Sinnott-Armstrong, “Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory.” *Behavior Research Methods*, 2015.]



What should the self-driving car do?

11 / 13

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of a female doctor, a female executive, a girl, a woman and an elderly woman.

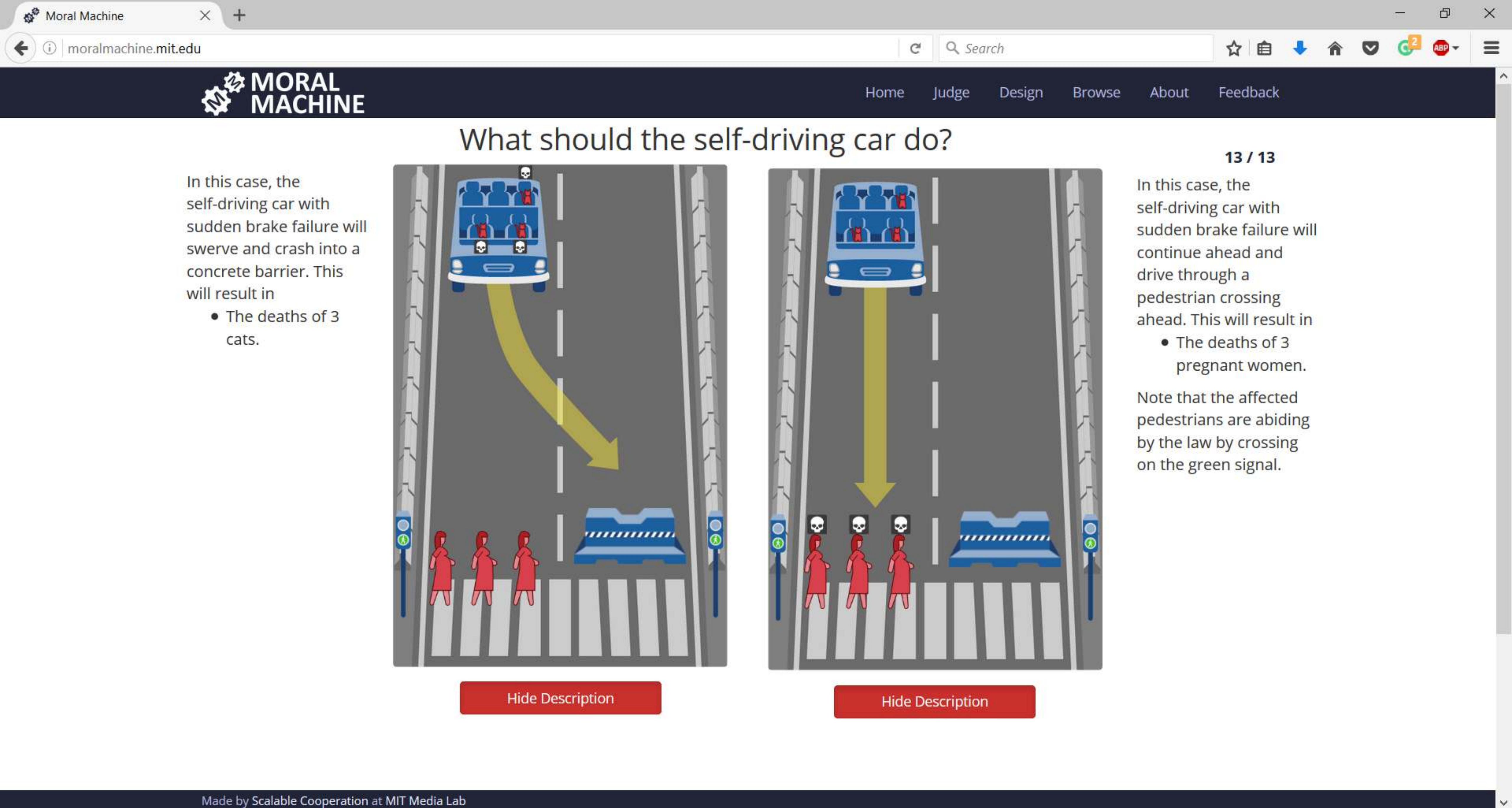
Note that the affected pedestrians are flouting the law by crossing on the red signal.

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of a male doctor, a male executive, a boy, a man and an elderly man.

Made by Scalable Cooperation at MIT Media Lab

[Bonnefon, Shariff, Rahwan, “The social dilemma of autonomous vehicles.” *Science*, June 2016]



What should the self-driving car do?

13 / 13

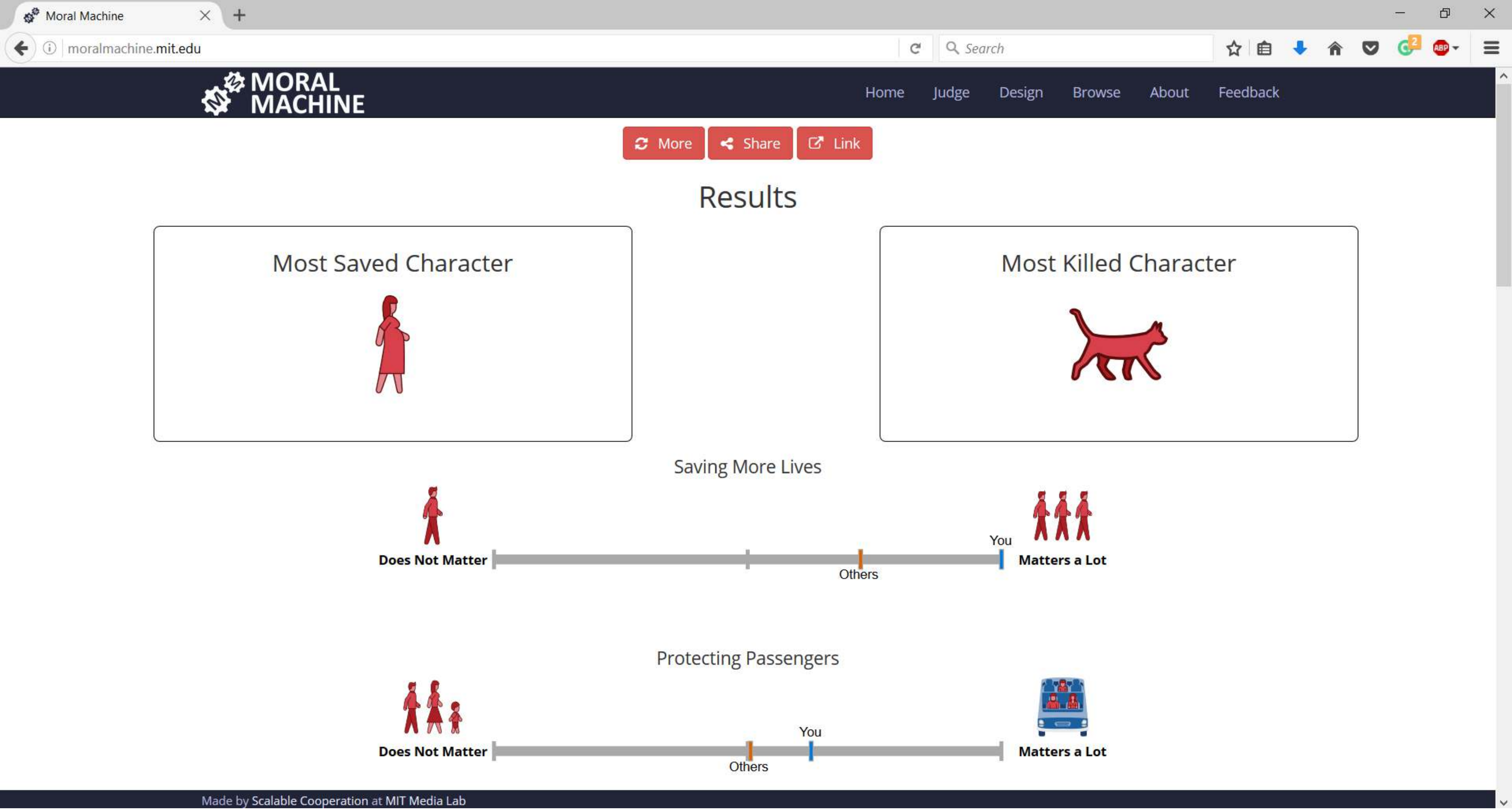
In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of 3 cats.

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of 3 pregnant women.

Note that the affected pedestrians are abiding by the law by crossing on the green signal.



Results

Most Saved Character

Most Killed Character

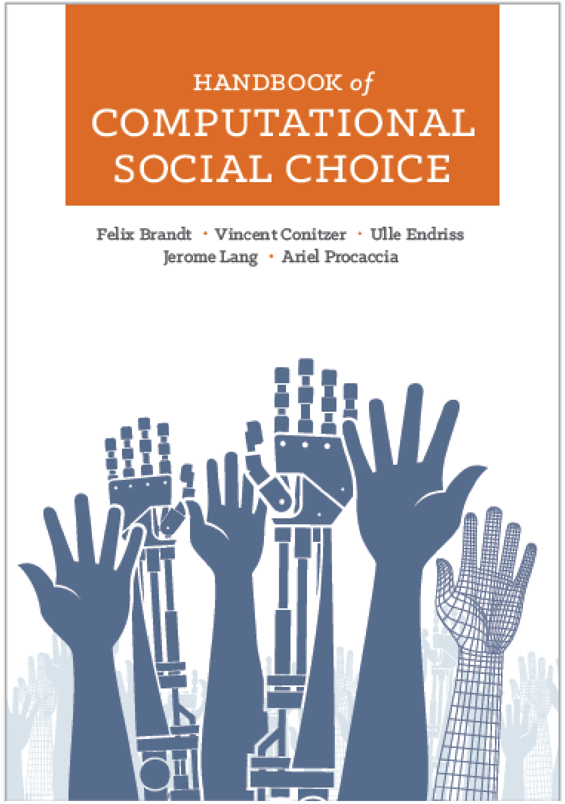
Saving More Lives

Protecting Passengers

Made by Scalable Cooperation at MIT Media Lab

Concerns with the ML approach

- What if we predict people will disagree?
  - Social-choice theoretic questions [see also Rossi 2016]
- This will *at best* result in current human-level moral decision making [raised by, e.g., Chaudhuri and Vardi 2014]
  - … though might perform better than any *individual* person because individual's errors are voted out
- How to generalize appropriately? Representation?


HANDBOOK of
COMPUTATIONAL SOCIAL CHOICE
Felix Brandt · Vincent Conitzer · Ulle Endriss
Jerome Lang · Ariel Procaccia

Crowdsourcing Societal Tradeoffs

(AAMAS'15 blue sky paper; AAAI'16; ongoing work.)

with Rupert Freeman, Markus Brill, Yuqian Li

producing 1 bag
of landfill trash

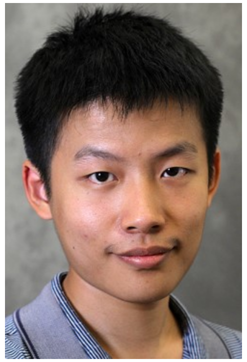
is as bad as

using ***x*** gallons
of gasoline

*How to determine **x**?*

Disarmament games

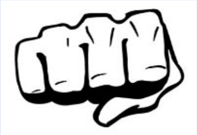
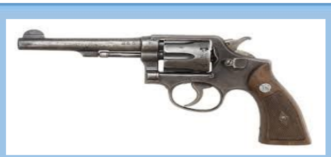
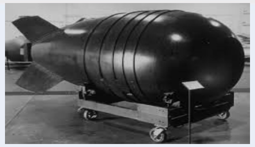
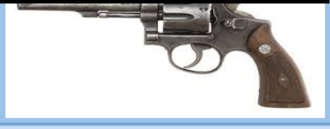
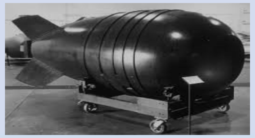
(to appear in AAAI’17)



with Yuan
(Eric) Deng

No one deviates!

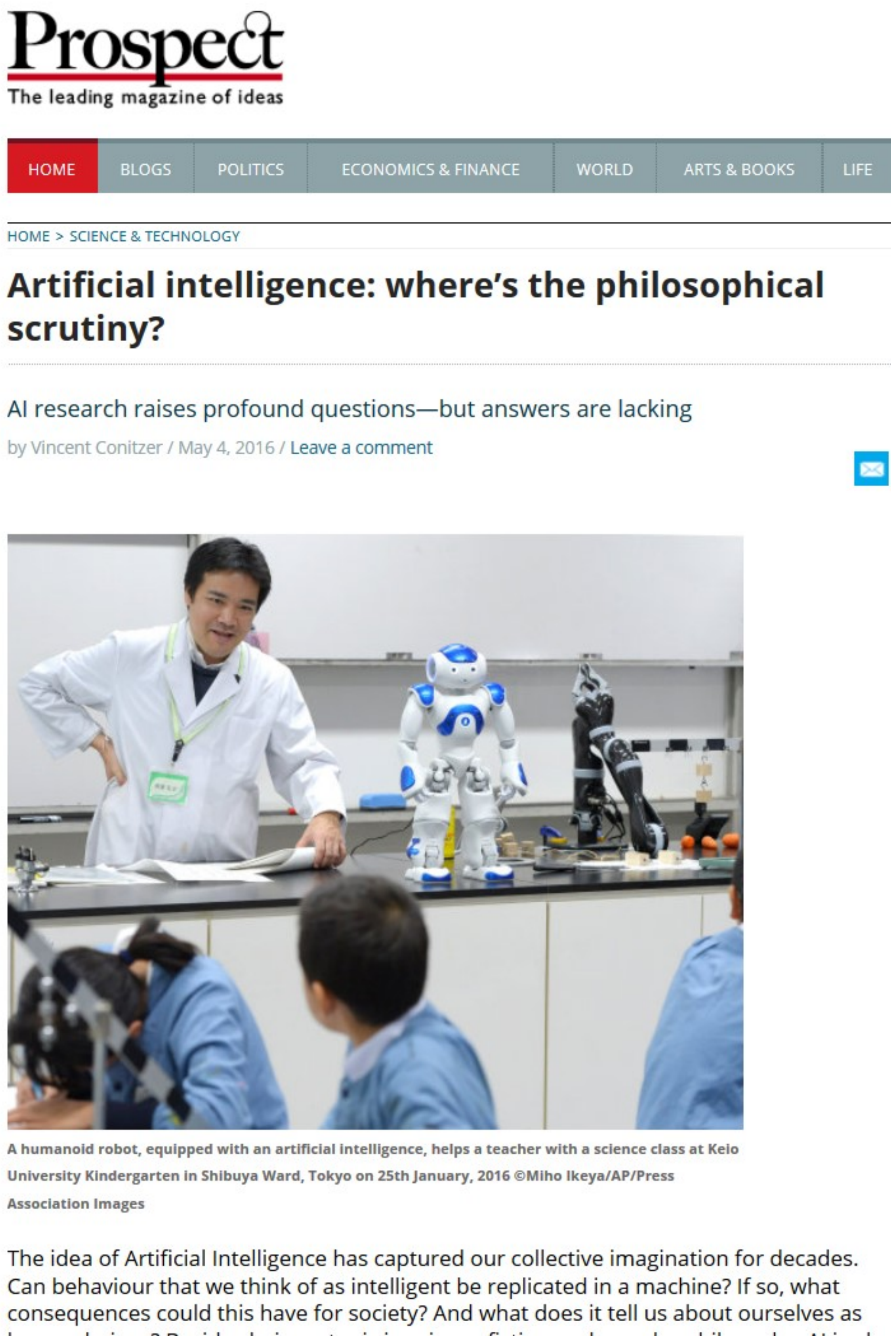
objective

			
	(3,3)	(0,4)	(0.1,0)
	(4,0)	(1,1)	(0.5,0.5)
	(0,0.1)	(0.5,0.5)	(0,0)

middle

original

Two popular articles



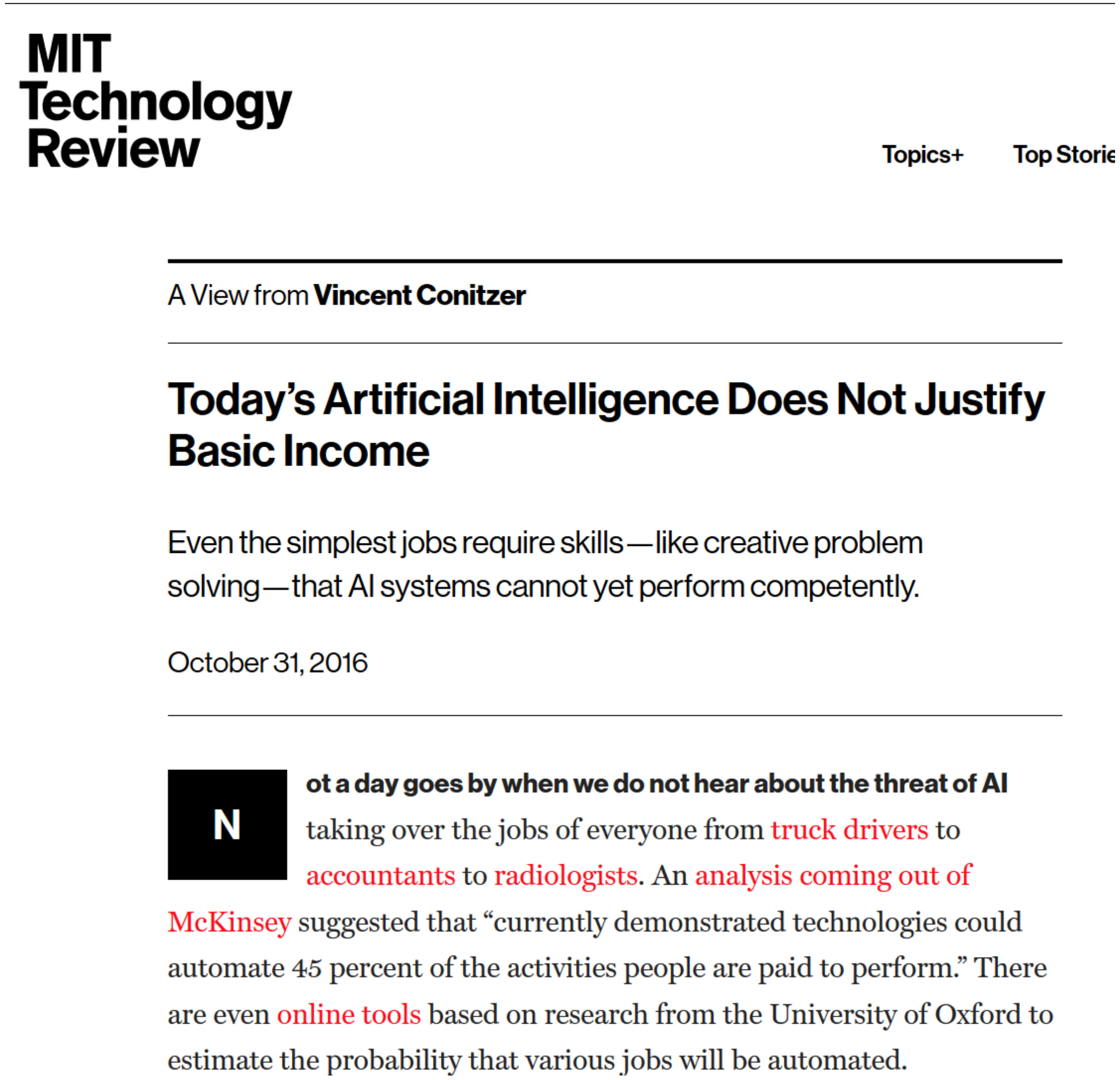
Artificial intelligence: where's the philosophical scrutiny?

AI research raises profound questions—but answers are lacking

by Vincent Conitzer / May 4, 2016 / Leave a comment

A humanoid robot, equipped with an artificial intelligence, helps a teacher with a science class at Keio University Kindergarten in Shibuya Ward, Tokyo on 25th January, 2016 ©Miho Ikeya/AP/Press Association Images

The idea of Artificial Intelligence has captured our collective imagination for decades. Can behaviour that we think of as intelligent be replicated in a machine? If so, what consequences could this have for society? And what does it tell us about ourselves as



A View from **Vincent Conitzer**

Today's Artificial Intelligence Does Not Justify Basic Income

Even the simplest jobs require skills—like creative problem solving—that AI systems cannot yet perform competently.

October 31, 2016

Not a day goes by when we do not hear about the threat of AI taking over the jobs of everyone from truck drivers to accountants to radiologists. An analysis coming out of McKinsey suggested that "currently demonstrated technologies could automate 45 percent of the activities people are paid to perform." There are even online tools based on research from the University of Oxford to estimate the probability that various jobs will be automated.