# Computer Science and Artificial Intelligence (P5)

# AI Concerns

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

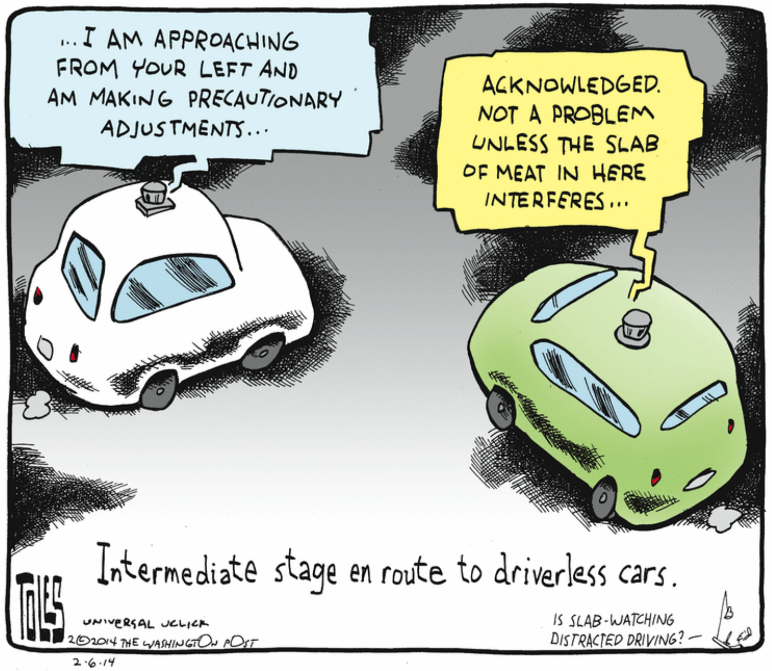
# AI ethics

- □ The world of AI is extremely **dynamic**.
- □ AI has the potential to solve problem faster, but it can also have **harmful consequences**
- □ How to use **AI responsibly** is hot topic in AI
- □ The five pillars of **AI ethics** are explainability, transparency, robustness, privacy and fairness.
- □ AI ethics must be considered throughout the entire AI lifecycle

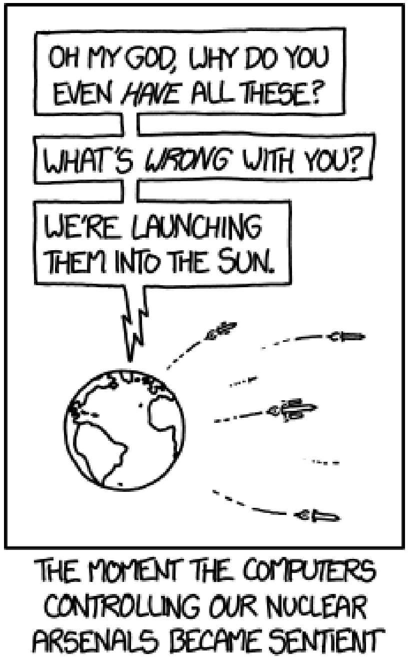
# Worries about AI - near term

SCHEDULE A MEETING WITH DILBERT AND ALICE FOR NEXT TUESDAY AT TEN.
DONE
Dilbert.com DilbertCartoonist@gmail.com
NEVER MIND. MY PHONE TOOK CARE OF IT.
12-16-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick
AWKWARD.

technological

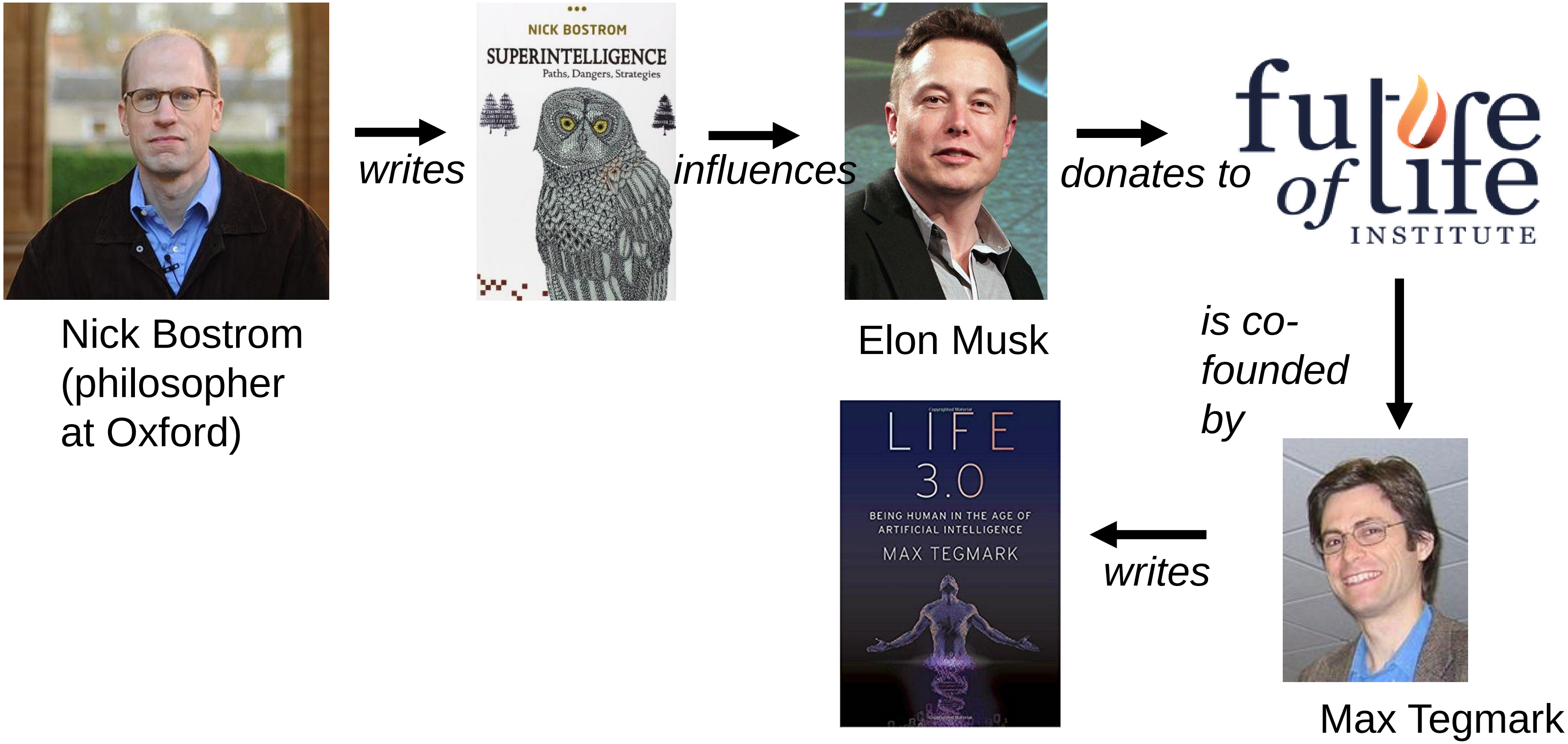

...I AM APPROACHING FROM YOUR LEFT AND AM MAKING PRECAUTIONARY ADJUSTMENTS...
ACKNOWLEDGED. NOT A PROBLEM UNLESS THE SLAB OF MEAT IN HERE INTERFERES...
Intermediate stage en route to driverless cars.
TOLES
UNIVERSAL UCLICK
2©2014 THE WASHINGTON POST
2·6·14
IS SLAB-WATCHING DISTRACTED DRIVING?

autonomous vehicles – legal and other issues


OH MY GOD, WHY DO YOU EVEN HAVE ALL THESE?
WHAT'S WRONG WITH YOU?
WE'RE LAUNCHING THEM INTO THE SUN.
THE MOMENT THE COMPUTERS CONTROLLING OUR NUCLEAR ARSENALS BECAME SENTIENT

autonomous weapon systems
…

# Worries about AI - superintelligence



Nick Bostrom (philosopher at Oxford)

*writes*

NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

*influences*

Elon Musk

*donates to*

*is co-founded by*

Max Tegmark

*writes*

LIFE 3.0
BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE
MAX TEGMARK

# AI concern example

- Face recognition
- Marketing
- Job candidate recommend

# What is Ethics

- □ **Ethics**: culturally agreed-upon morals and standards of action
- □ Ethics guide their decision-making, especially for decisions that impact others

# AI Ethics

□ **AI must be built with ethics** at the core so its outcomes *align with human ethics* and expectations

□ AI ethics is a multidisciplinary field that investigates how to **maximize AI's beneficial impacts** while **reducing risks** and **adverse impacts**.

# AI Principles

- The purpose of AI is to augment human intelligence
  - Do not seek to replace human intelligence with AI, but support it
- Data and insights belong to their creator
  - Company has not and will not provide government access to client data for any surveillance programs, and it remains committed to protecting the privacy of its clients.
- AI systems must be transparent and explainable
  - Companies must be clear about who trains their AI systems, what data was used in training and, most importantly, what went into their algorithms' recommendations.

# AI Ethics

- Five pillars for AI ethics: explainability, fairness, robustness, transparency, and privacy.
- These pillars are focus areas that help us take action to build and use AI ethically.

# Explainability

□ AI is **explainable** when it can show **how** and **why** it arrived at a particular outcome or recommendation.

□ Explainability: an AI system showing its work.

# Fairness

- □ AI is **fair** when it treats individuals or groups *equitably*.
- □ AI can help humans make fairer choices by counterbalancing human biases, but beware — bias can be present in AI too, so steps must be taken to mitigate it.

# Robustness

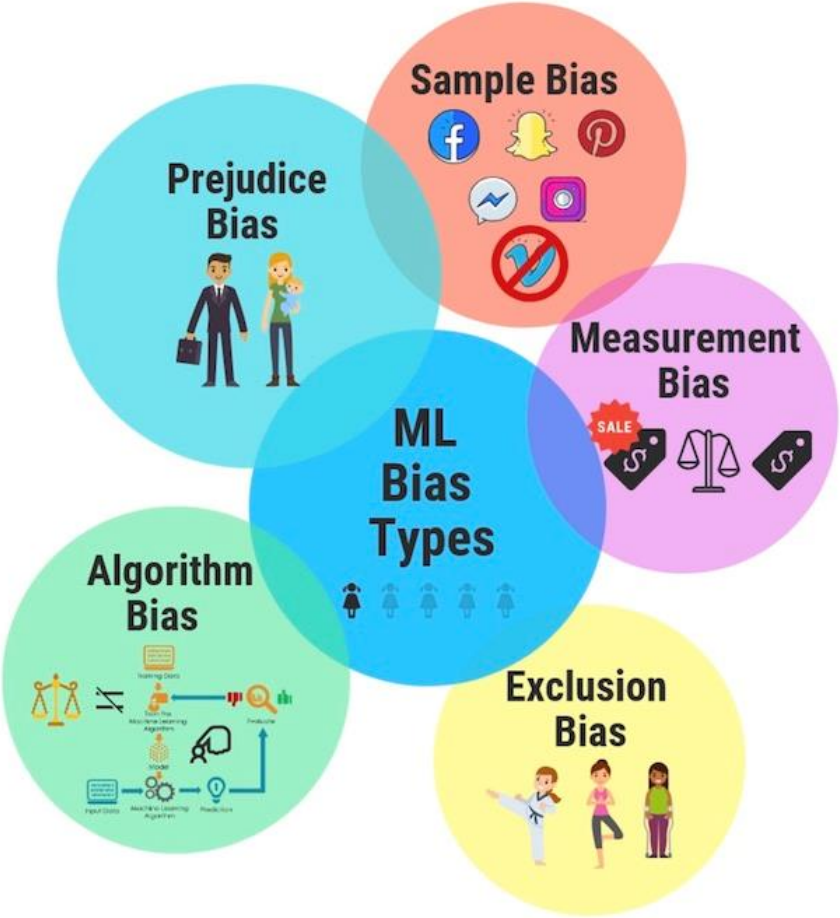
- □ AI is **robust** when it can effectively handle exceptional conditions, like abnormal input or adversarial attacks.
- □ Robust AI is built to withstand intentional and unintentional interference.

# Transparency

- □ AI is **transparent** when appropriate information is shared with humans about how the AI system was designed and developed.
- □ Humans have access to information like what data was used to train the AI system, how the system collects and stores data, and who has access to the data the system collects.

# Privacy

- □ Because AI ingests so much data, it must be designed to prioritize and safeguard **humans' privacy and data rights**.
- □ AI that is built to respect privacy collects and stores only the **minimum** amount of data it needs to function, and collected data should never be repurposed **without users' consent**, among other considerations.


Sample Bias
Prejudice Bias
Measurement Bias
SALE
ML Bias Types
Algorithm Bias
Training Data
Train the Machine Learning Algorithm
Feedback
Model
Input Data
Machine Learning Algorithm
Prediction
Exclusion Bias

# AI BIAS

# AI bias

- □ What is bias in the context of AI.
- □ How bias can emerge in AI.
- □ How bias can impact AI's outcomes
- □ How to begin mitigating potential bias in AI.

# What is bias in AI?

□ Bias is all about unwanted behaviors.

□ Bias gives systematic disadvantages to certain groups and individuals.

# How can bias emerge in AI?

- AI is trained on historical decisions that human decision makers have made in the past.
  - The human decision makers in the past were implicitly or explicitly biased themselves and so that's reflected in the training data through prejudice.
- The sampling of the data.
  - Certain groups are overrepresented or underrepresented in a particular data set.
  - The data processing or data preparation phase in the data science project.
- The way that the problem is conceptualized can also introduced bias.

# How to mitigating bias in AI.

- Having a diversed team of people.
  - Help recognizing that there are biases
- Search for high-quality data sets
- There are many possible technical approaches to mitigate bias.
  - Some ML model that help reduce bias

# Quiz 1

□ What is AI ethics?

A. How and why an AI system arrived at a particular outcome or recommendation

B. How to build and use AI in ways that align with human ethics and expectations

C. An organization’s act of governing AI through its corporate instructions, staff, processes, and systems

D. A multidisciplinary field that investigates how to maximize AI’s beneficial impacts while reducing risks and adverse impacts

# Quiz 2

□ What are the pillars of AI ethics?

A. Environmental impact, equitable impact, ethical impact

B. Explainability, fairness, robustness, transparency, privacy

C. Trust, efficiency, compliance

D. Awareness, governance, operationalization

# Quiz 3

□ In AI, what is explainability?

A. An AI system’s ability to effectively handle exceptional conditions, like abnormal input or adversarial attacks

B. An AI system’s ability to show how and why it arrived at a particular outcome or recommendation

C. An AI system’s ability to treat individuals or groups equitably

D. When appropriate information is shared with humans about how an AI system was designed and developed

# Quiz 4

□ In AI, what is robustness?

A. An AI system’s ability to treat individuals or groups equitably
B. An AI system’s ability to prioritize and safeguard humans’ privacy and data rights
C. An AI system’s ability to show how and why it arrived at a particular outcome or recommendation
D. An AI system’s ability to effectively handle exceptional conditions, like abnormal input or adversarial attacks

# Quiz 5

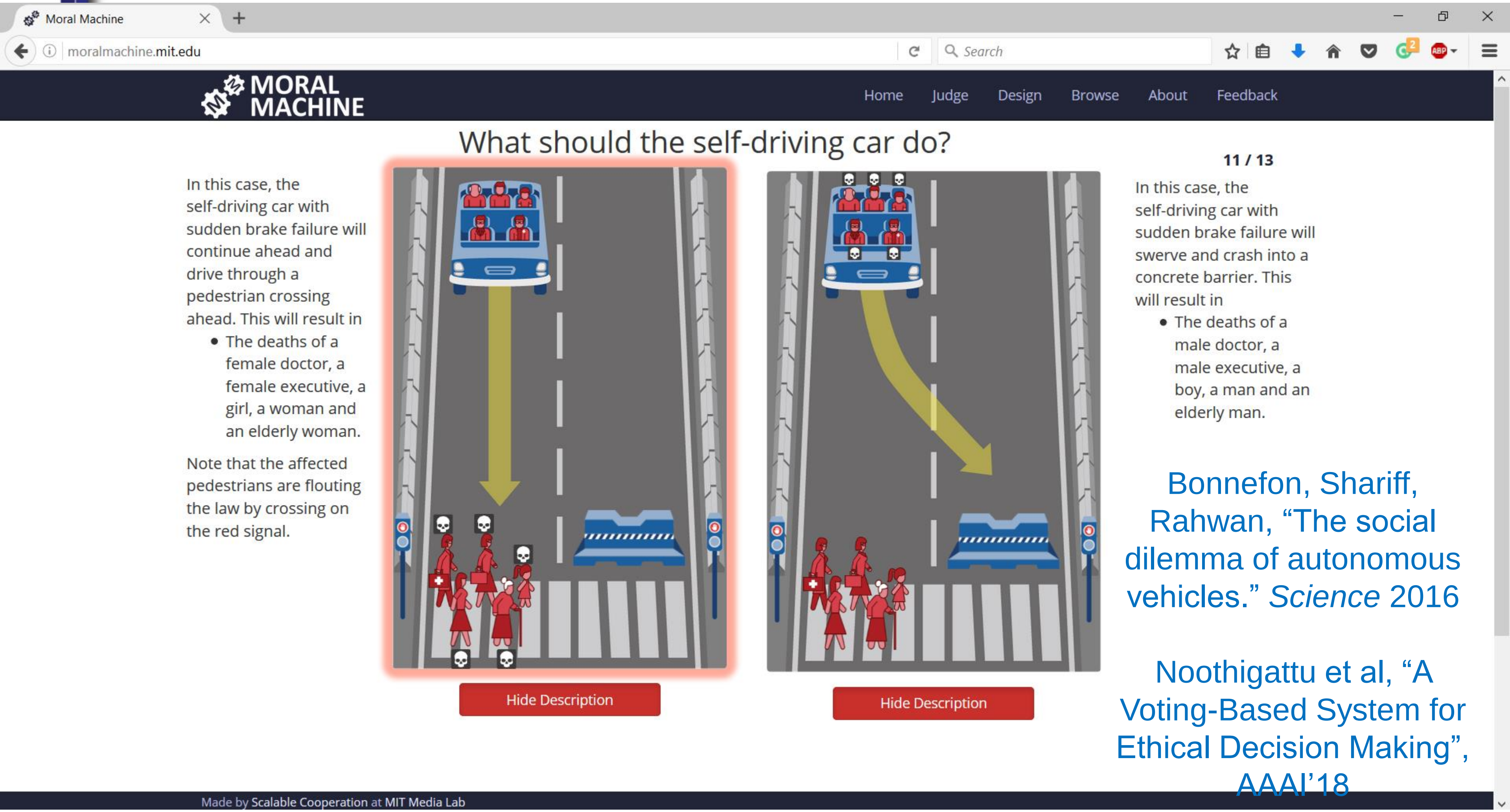
□ In AI, what does bias do?

A. Augments human intelligence

B. Gives systematic disadvantages to certain groups or individuals

C. Identifies and addresses socio-technical issues raised by AI

D. Solves problems faster



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.

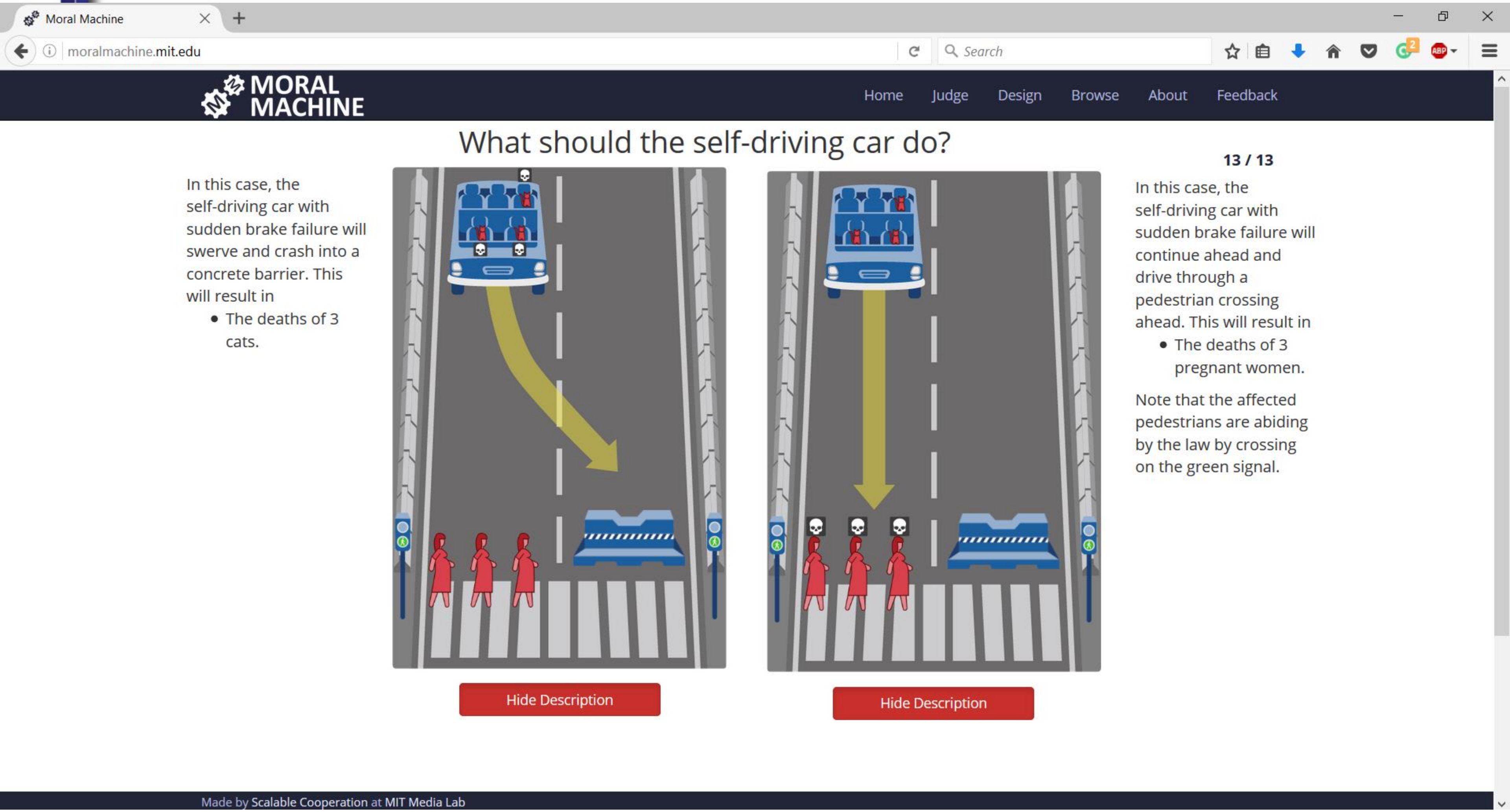
Note that the affected pedestrians are flouting the law by crossing on the red signal.

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.

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

Noothigattu et al, “A Voting-Based System for Ethical Decision Making”, AAAI’18



## What should the self-driving car do?

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.

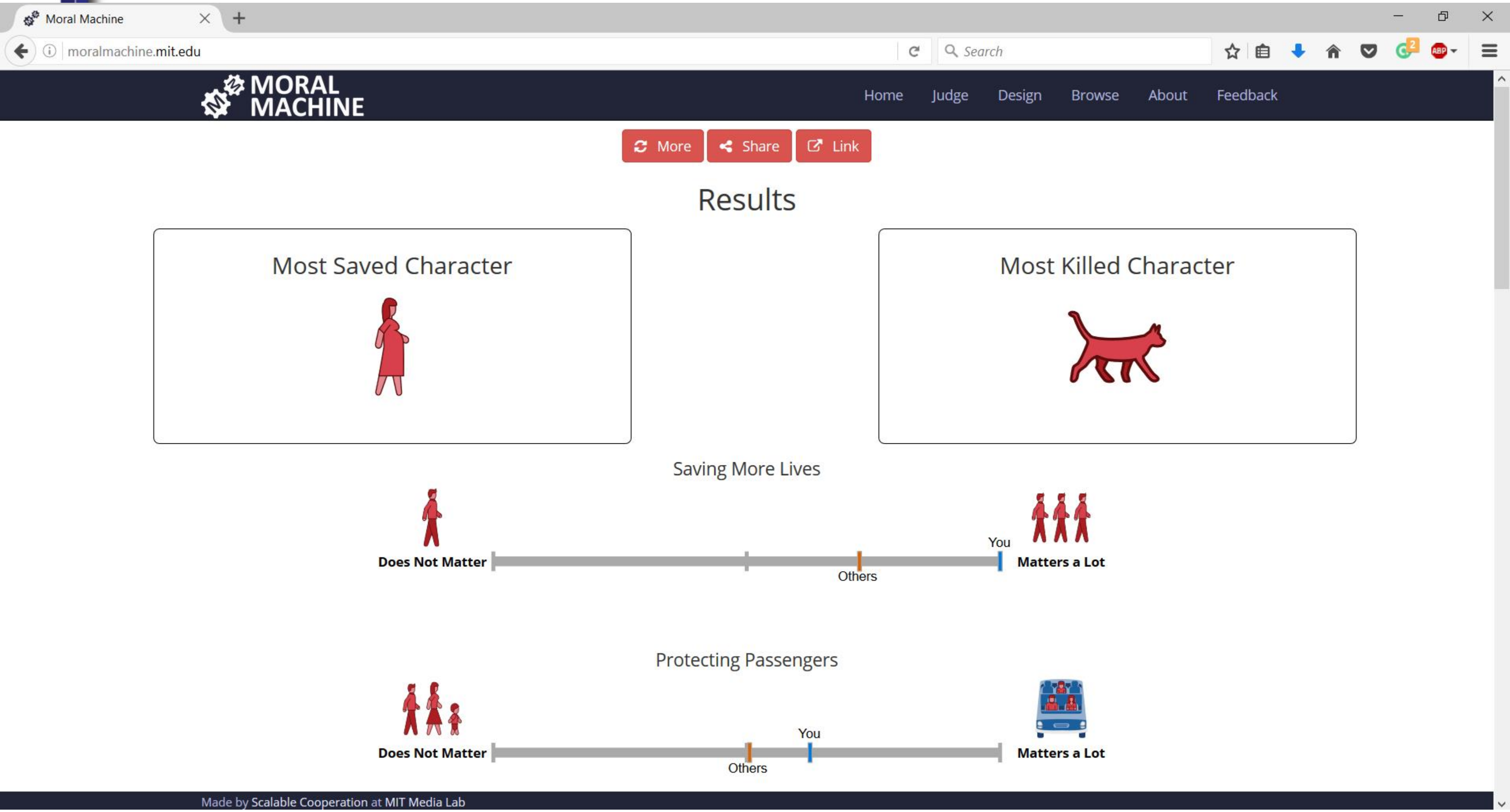
13 / 13

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 3 pregnant women.

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

Made by Scalable Cooperation at MIT Media Lab



# 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, and Noothigattu et al. 2018 for moral machine data]
- 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?

# Reference

- □ [1] Introduction to AI course, coursera
- □ as