THE UNIVERSITY OF
MELBOURNE

13. Ethics & autonomous agents

—

Adrian Pearce
School of Computing and Information Systems
University of Melbourne

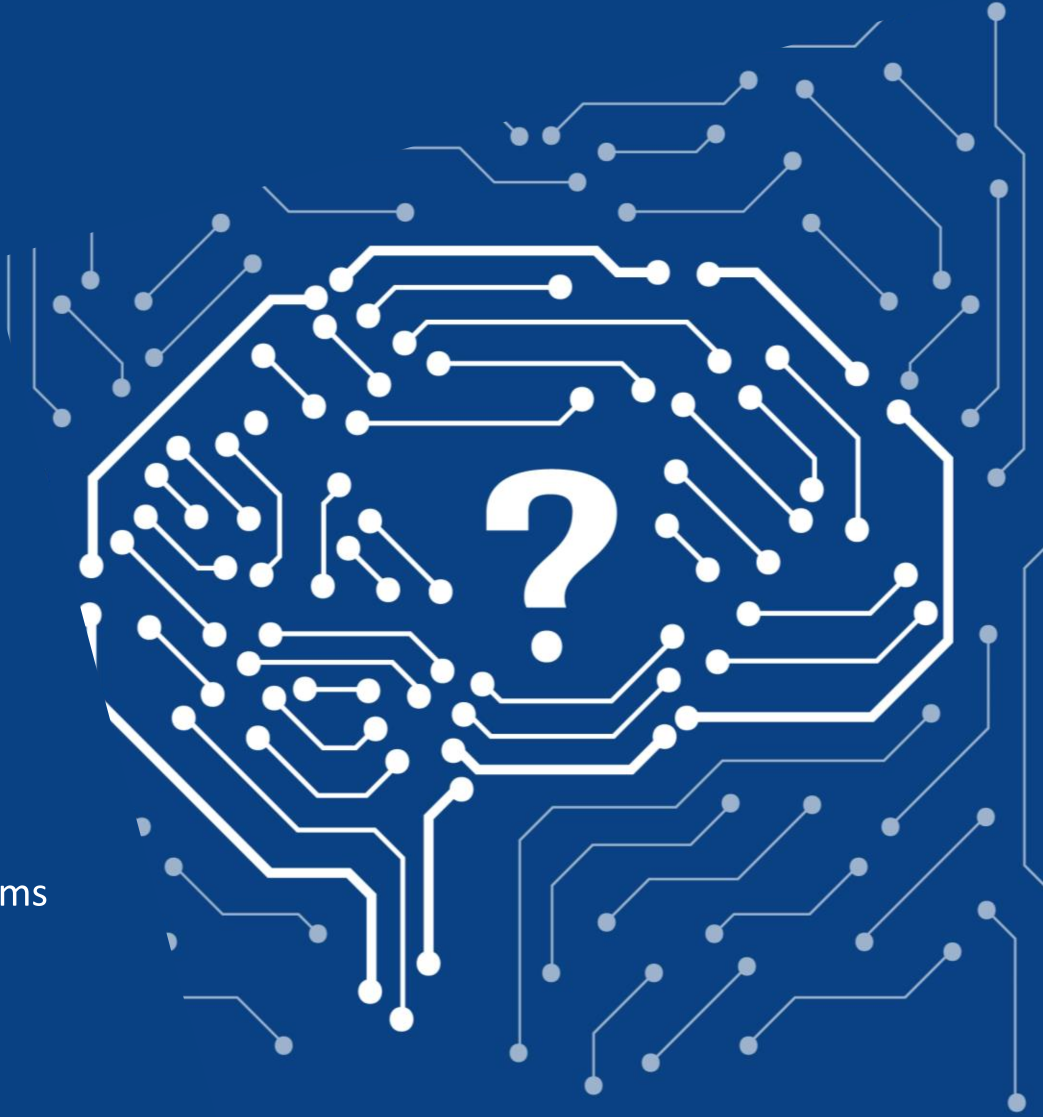


Image source: https://www.google.com/selfdrivingcar/

The Trolley Problem

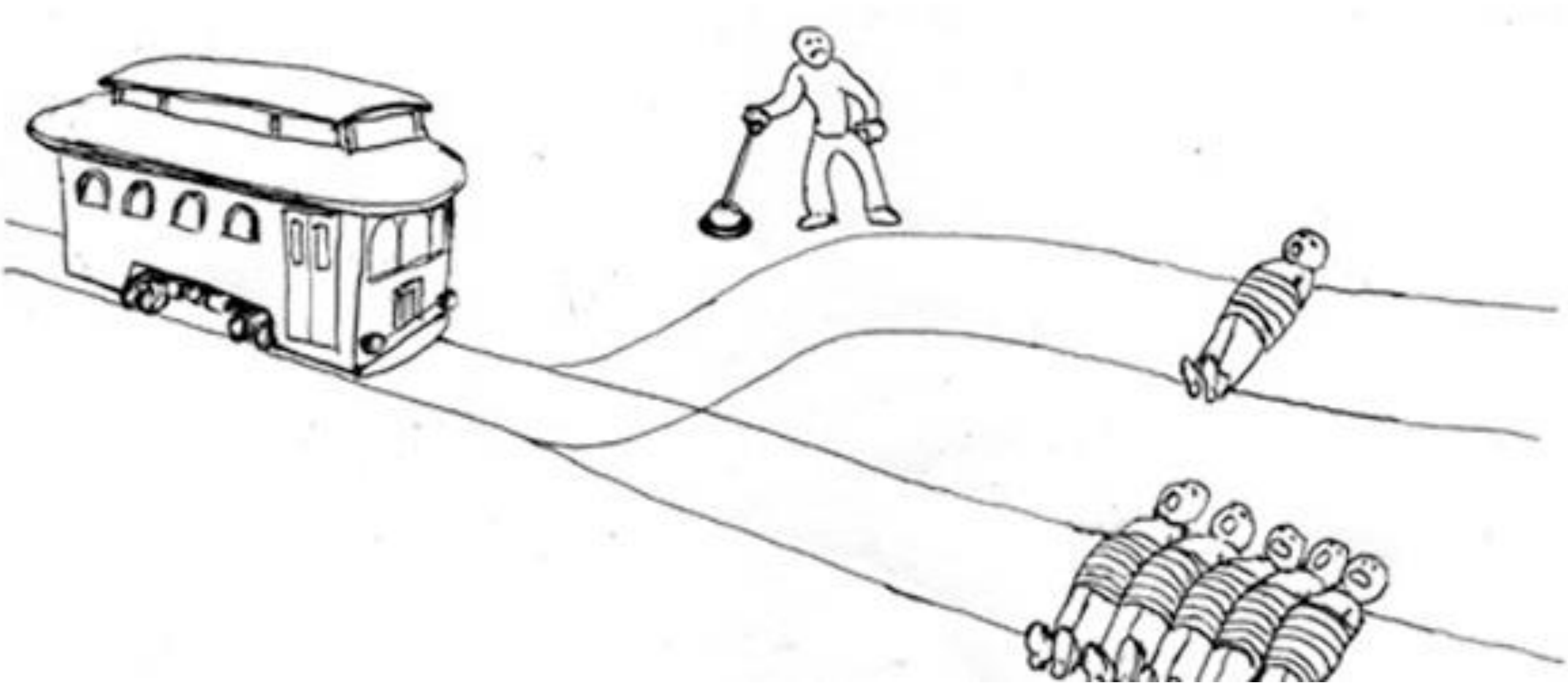


Image source: http://www.relativelyinteresting.com/the-trolley-problem-a-thought-experiment-that-tests-our-morality/

Ethics

Utilitarianistic ethics: only outcomes matter, and we should always opt for the best.

Deontological ethics: actions matter too! We have a duty to help others as well as a duty not to harm others, but the latter is stronger than the former.

What should we do?

The ‘Fat Man’ Trolley Problem

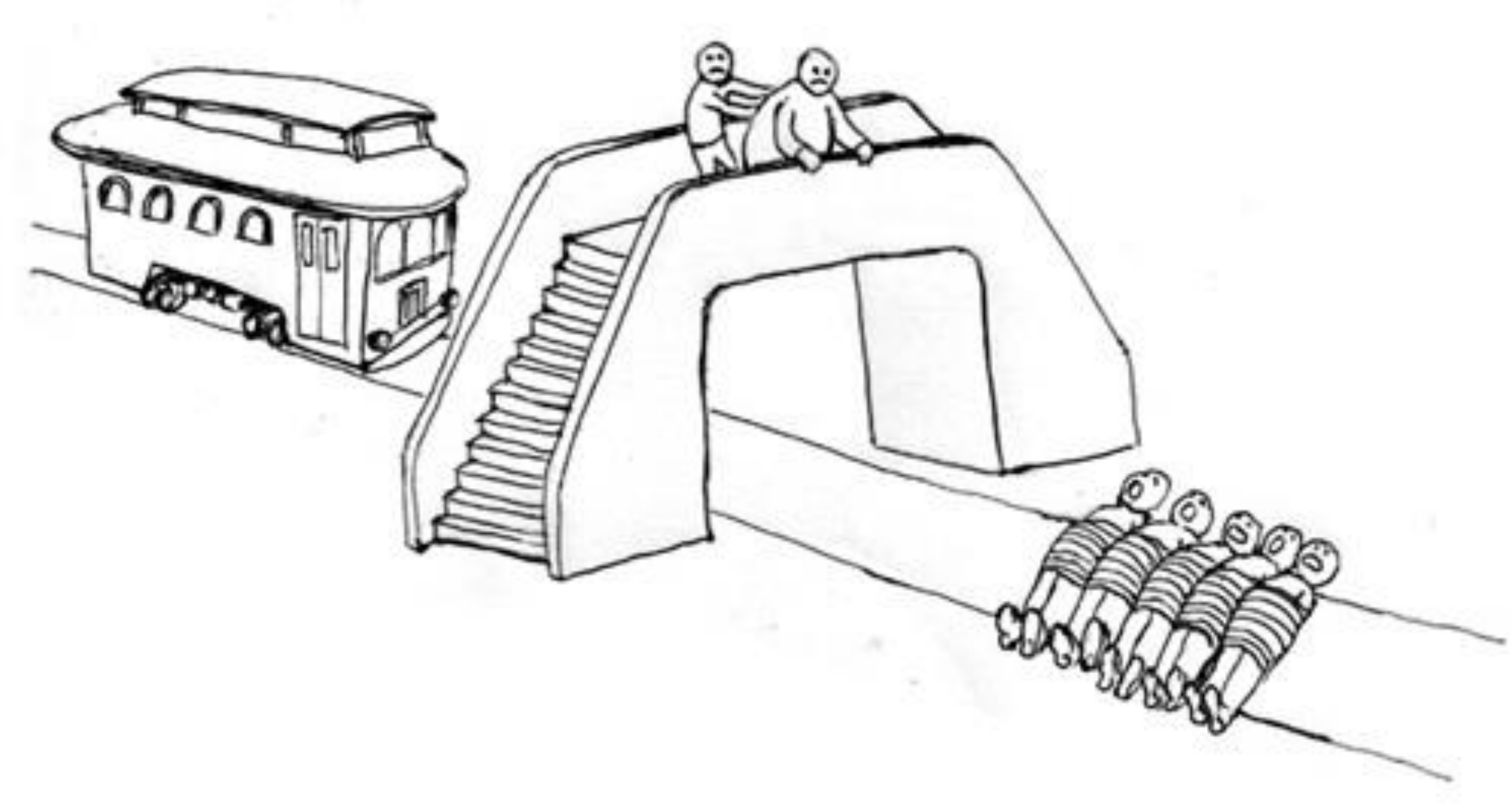
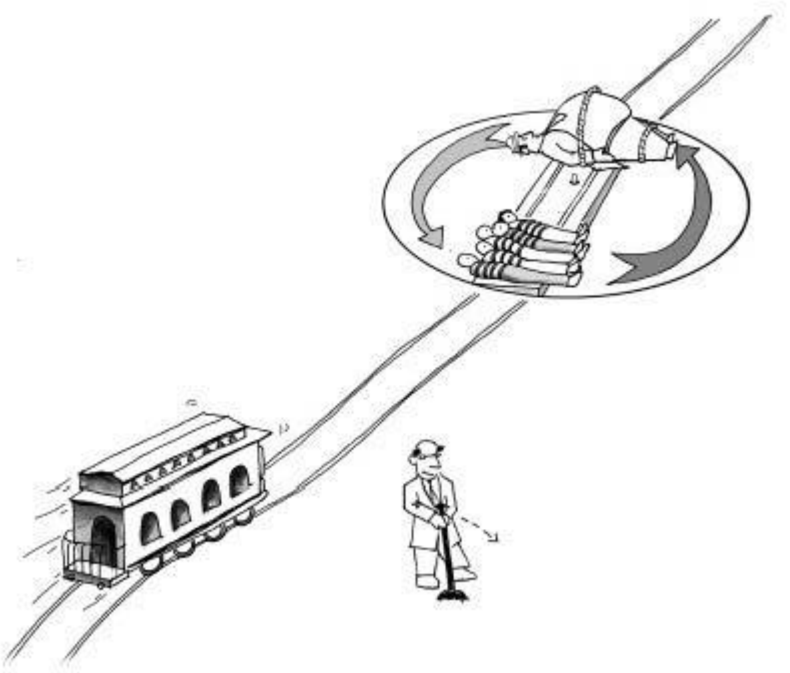


Image source: http://www.relativelyinteresting.com/the-trolley-problem-a-thought-experiment-that-tests-our-morality/

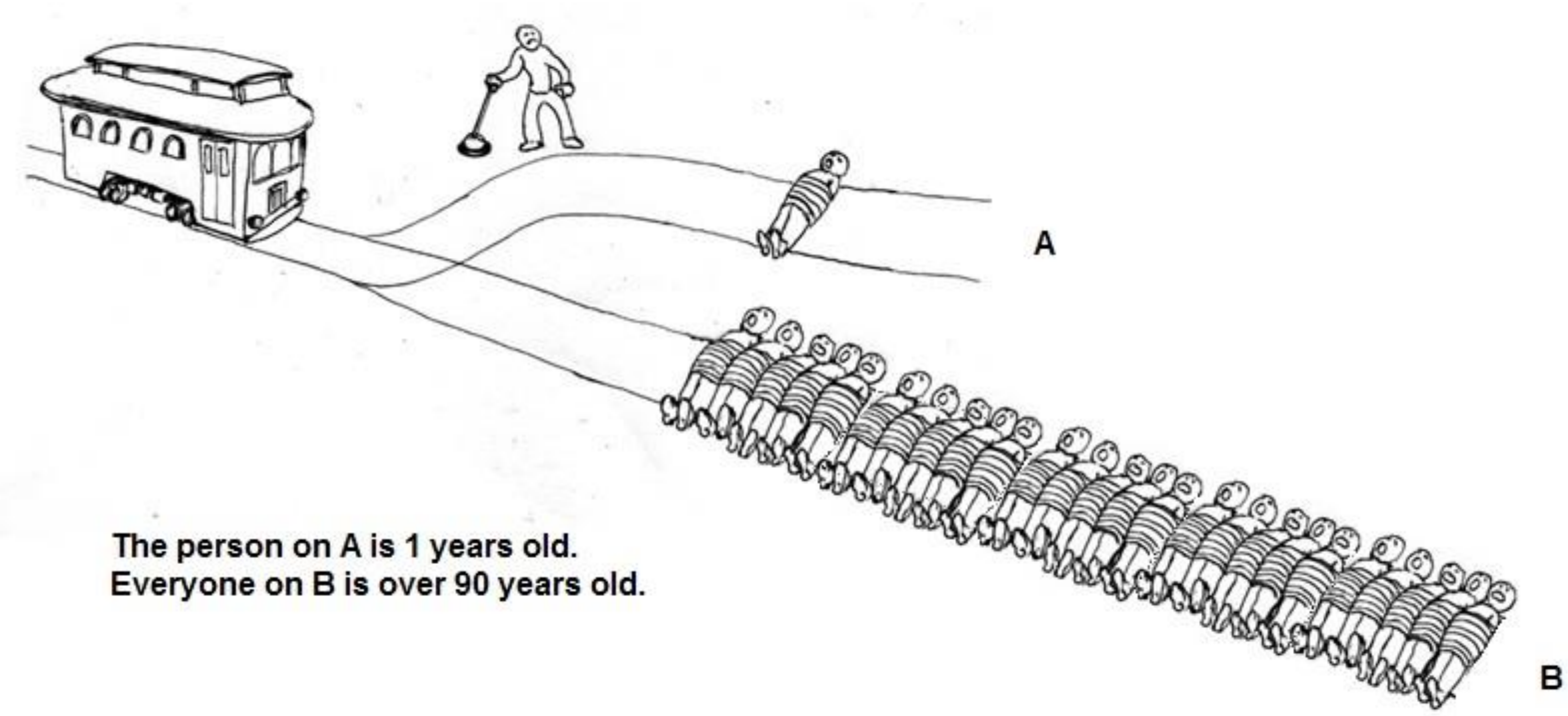
The ‘Fat Man’ Trolley Problem (Lazy Susan)



In Lazy Susan you can save the five by revolving the turntable 180 degrees—this will have the unfortunate consequence of placing one man directly in the path of the train. Should you rotate the Lazy Susan?

(Figure: Would you kill the Fat Man?)

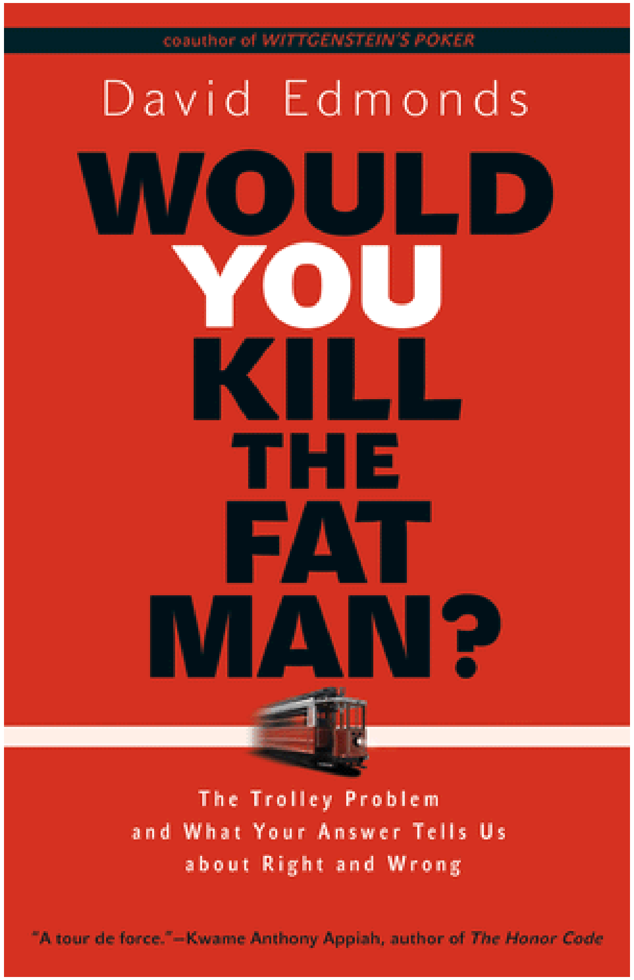
Many variants of the Trolley Problem



The person on A is 1 years old.
Everyone on B is over 90 years old.

Image source: http://www.relativelyinteresting.com/the-trolley-problem-a-thought-experiment-that-tests-our-morality/

Experimental ethics



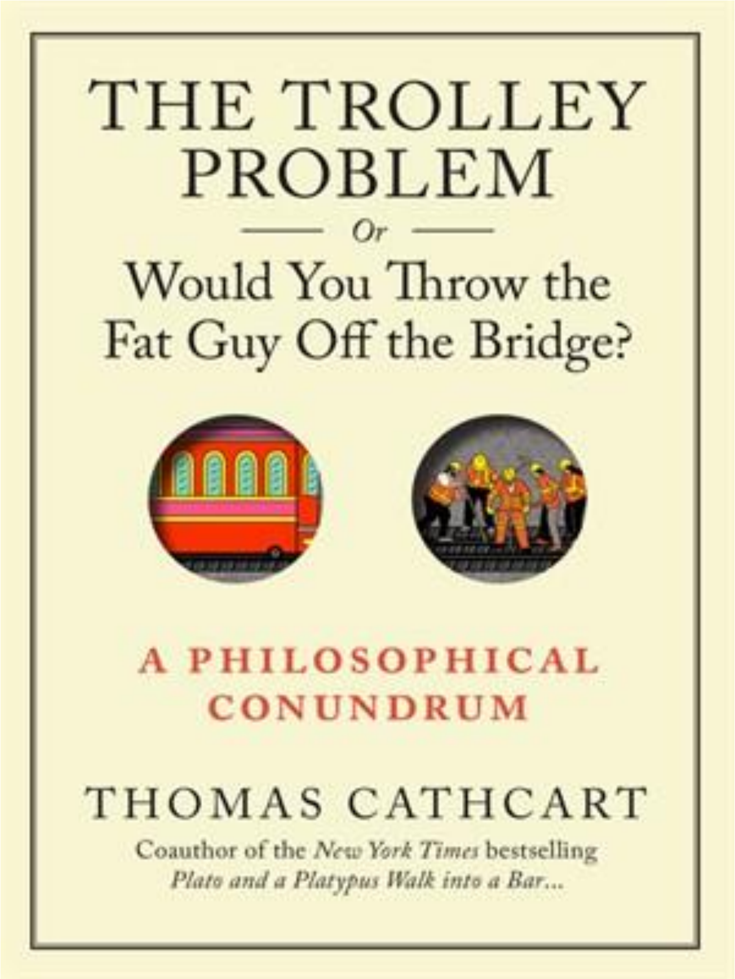
coauthor of *WITTGENSTEIN'S POKER*

David Edmonds

WOULD YOU KILL THE FAT MAN?

The Trolley Problem and What Your Answer Tells Us about Right and Wrong

"A tour de force."—Kwame Anthony Appiah, author of *The Honor Code*



THE TROLLEY PROBLEM

Or

Would You Throw the Fat Guy Off the Bridge?

A PHILOSOPHICAL CONUNDRUM

THOMAS CATHCART

Coauthor of the *New York Times* bestselling *Plato and a Platypus Walk into a Bar...*

Pre-programmed ethics: The case of autonomous vehicles

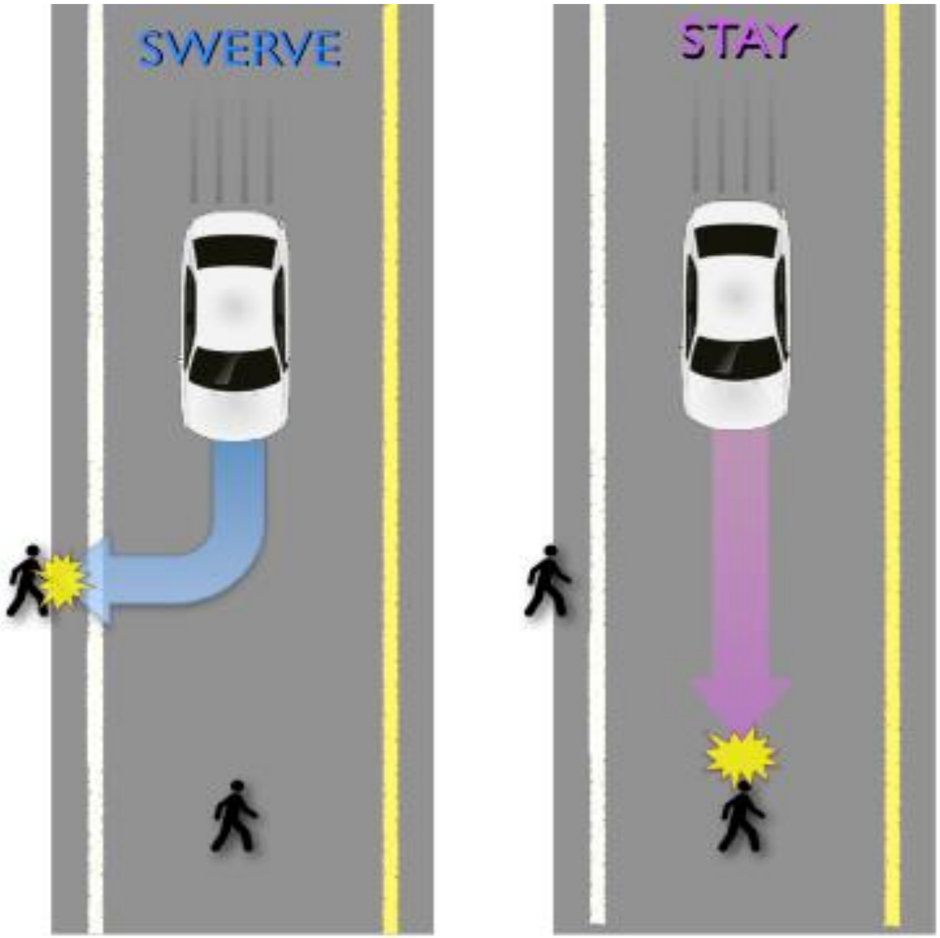
Split-second driver decision:

- reckless driver → criminal liability
- otherwise → blame “fate”, “bad luck”, “heat of the moment”, and do not blame the driver

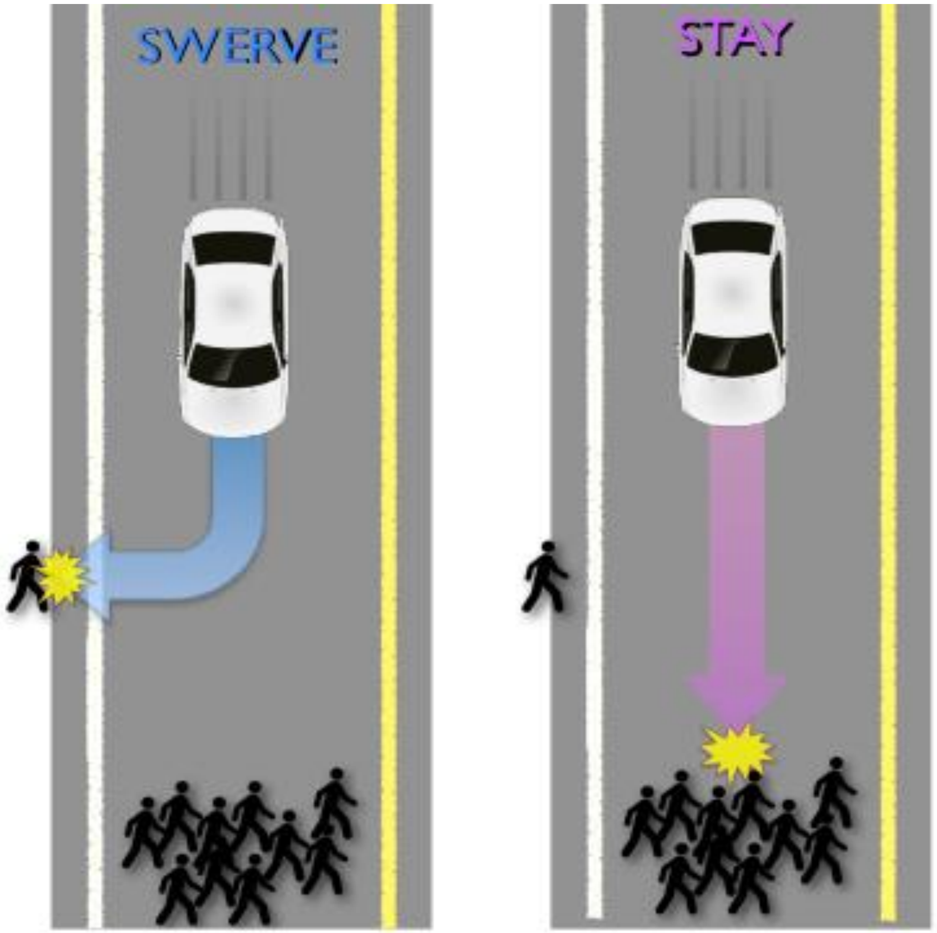
Automated decision-making:

- We must decide in advance! No split-second decisions
- What behaviour should be programmed?

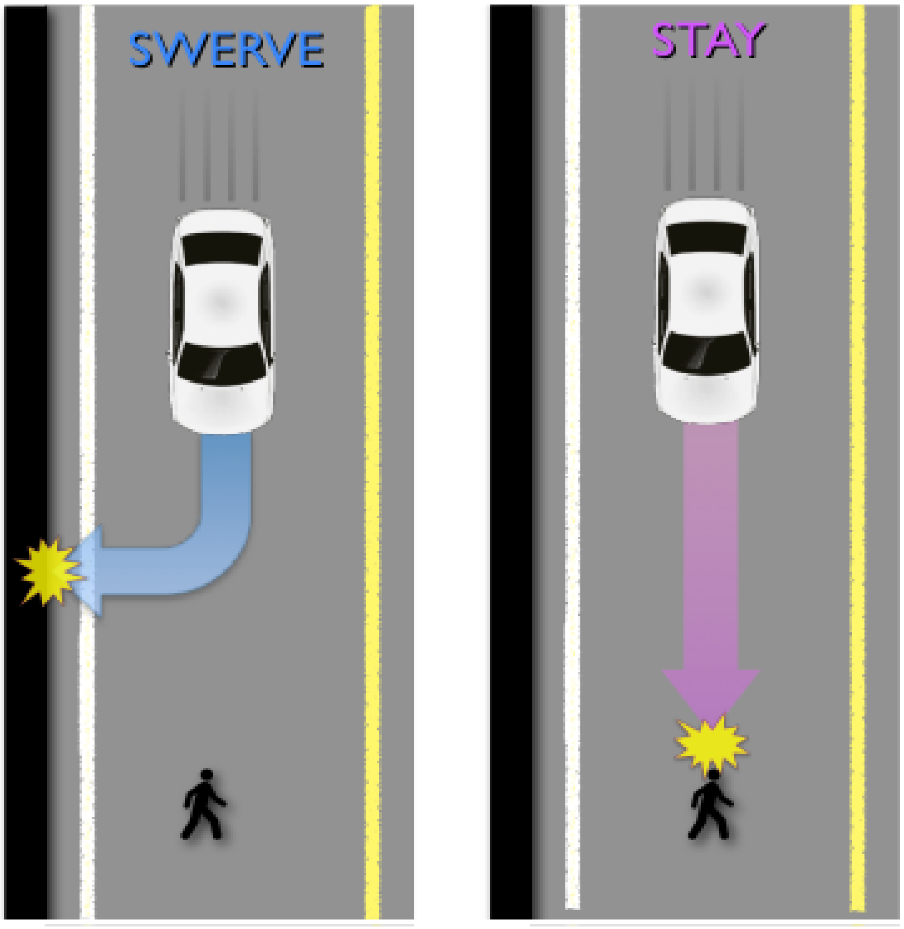
Swerve or stay?



Swerve or stay?



Swerve or stay?



In some larger-scale experiments

Would you buy a self-driving car that protected passengers instead of the driver?

Despite preferring self-sacrificing vehicles ...

… respondents were less likely to buy such cars.

Trade-offs must be made by law-makers to solve this *social dilemma*

But.....

What about related but different scenarios:

- 99.95% confidence an object is not a person
- What if the cars knows there is a child on board?
- Buyer liability: If buyers can choose the “moral algorithm” are they liable for its choices?

Fairness

In **data mining** and **computer vision**, a machine learning algorithm is said to be fair, or to have fairness, if its results are independent of given variables, especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome

- For example gender, ethnicity, etc. (there are well-known examples of unfair ML implementations, e.g. law enforcement and recruitment)

Here is a simple (although not machine learning) example demonstrating why we must take care of data distributions when developing (any) algorithms:

- In 2020, UK dropped its (COVID-19 motivated) exam-grading policy for General Certificate of Secondary Education (GCSE) results. You can watch this explainer (posted by the Financial Times) **What went wrong with the A-level algorithm?**
  - https://www.youtube.com/watch?v=jHtMLEhDOVE

Fairness in AI planning & reinforcement learning

What about learning behaviour for augmented (human-agent) intelligence or fully autonomous systems?

- Model-free AI planning and reinforcement learning **is potentially subject to the same fairness limitations** as in computer vision and data mining, due to for example
  - sampling complexity (ability to sample enough data and/or query simulators to generate meaningful probability distributions)
  - non-stationarity (when the future is not symmetrical with the past independent of the amount of data sampled – adversarial and competitive games are typical examples)
- model-based approaches rely on the correctness/completeness of the *modelling language* encoding

Who should decide?

AI programmers?

Companies?

Lawmakers?

The public?

Ethicists and philosophers?

The Moral Machine at MIT Media Lab

http://moralmachine.mit.edu/

Reading

(Chapter 5: Fat man, Loop and Lazy Susan) *Would you kill the fat man?* by David Edmonds, 2013

https://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=1275331

(External link, UniMelb login required to access e-book)

THE UNIVERSITY OF
MELBOURNE

Thank you