6CCS3AIN AI Reasoning & Decision-Making
AI and Ethics
Week 10

Peter McBurney
Department of Informatics
King's College London
London

peter.mcburney@kcl.ac.uk

Week 10
7 December 2020

AI and Ethics Outline

- This week we talk about AI and Ethics
- Why is this topic important
- Regulatory and legal constraints
- Ethical dimensions
  - Fairness (non-bias), transparency, explanation, rectification
- Features of this domain
- Deciding what we ourselves should do in specific situations.

Why is it important to consider ethics?

- Most technologies have good and evil applications
- As engineers we owe a duty to our society to consider the ethics of our work
  - eg, British Computer Society Code of Conduct
- With AI, there are particular aspects we need to consider
  - Algorithms may be learnt, so that even the software developers do not know what they do or how.
  - Many machine learning methods are “black boxes”
  - Data may be biased
  - There may be significant legal consequences to our design decisions.

Trolley Problems

- Thought experiments in which we are faced with a moral dilemma
  - Often involve the control of a trolley (ie, a tram)

- For example:
  - We are driving a car in the left land and we see a pedestrian in our lane in front of us
    - If we keep driving we will likely kill the pedestrian

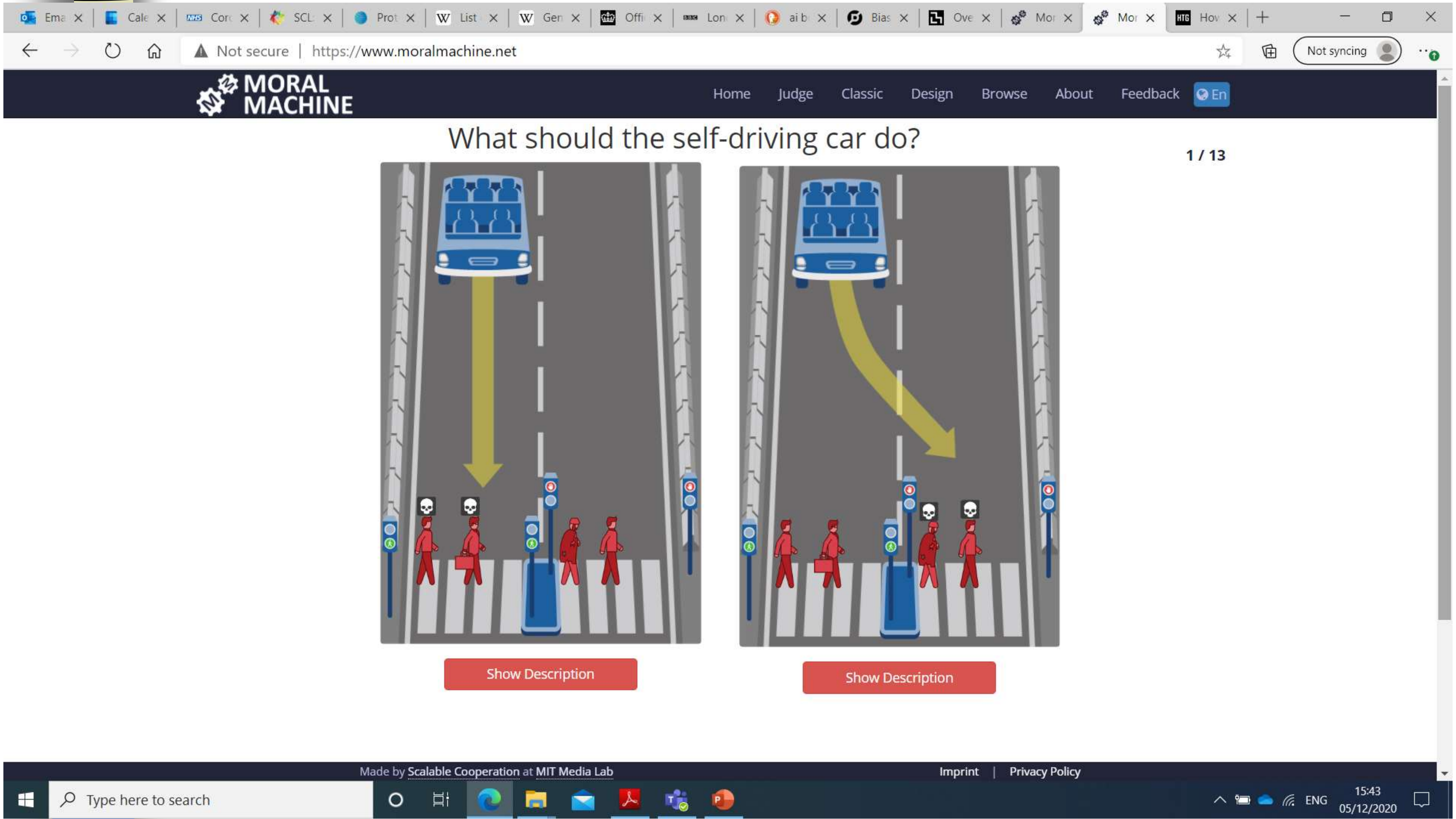
  - OR, we can swerve to the right lane, where there is a car heading towards us
    - If we swerve to the right lane, we will hit the car and this may kill us and the people in the car

  - What should we do?

  - Would our answer be different if the numbers of people impacted were different?

MIT Moral Machine Experiment

www.moralmachine.net



Aside – Norms vs Rules

- As AI engineers, we could hard-code many rules
  - Eg, the road rules that say we should always drive on the left
  - Hard-coding would mean the self-driving vehicle could NEVER drive on the right

- But sometimes we need the car to use the other lane, even though it is against the law
  - eg, Trolley Problems

- So, our usual solution is not to hard-code the rules as unbreakable constraints, but to code them as norms
  - Norm: an accepted standard or way of behaving, which most people follow
  - How should a machine know when it should break a norm?

- Norms and their exceptions are studied extensively in AI
  - Particularly in AI and Law.

Decisions often involve ethical trade-offs

- Trade-offs:

  Lives lost as a result of one action-option
  vs.
  Lives lost as a result of another action-option

- Maybe also another trade-off:

  Lives lost outside the car as a result of one action-option
  vs.
  Lives lost inside the car as a result of another action-option

- What car would you purchase?
- What car would you design?

Pressures from regulators

- Our Society is very concerned with various aspects of new technologies, such as AI

- AI is an important current focus of regulators, eg
  - European Commission
    - New regulations on AI coming in 2021
    - GDPR (General Data Protection Regulations)
    - MiFID2 (Markets in Financial Instruments Directive 2)

  - UK likely to follow EU regulations for several years
    - UK Government Office for AI

  - National regulators for data privacy and human rights, eg,
    - UK Information Commissioners Office (ICO)
    - Singapore Personal Data Protection Commission
    - Australian Human Rights Commission (looking at rights under CCTV).

Regulatory focus has been on

- Fairness (and elimination of bias)
  - Systems should not be biased against particular groups
  - People with protected characteristics (age, gender, religion, ethnicity, etc)
- Transparency
  - Stakeholders should be able to see what input data is used, what processes or algorithms are used, what output data results, and what the intended and realized purposes are
- Explainability
  - Automated decision-making systems should be able to explain their decisions in a way that humans can understand
- Rectification
  - Automated decisions should be able to be reversed
- Human involvement
  - Are decisions mediated by humans in the loop
- Governance of AI systems
  - Singapore Government Personal Data Protection Commission (**PDPC**) Model AI Governance Framework (Second Edition), released January 2020.

Legal Aspects

How can responsibility be attributed?

AI has no separate legal personality and cannot be an inventor for patents

– *Stephen L Thaler v Comptroller-General of Patents, Designs and Trade Marks* [2020] EWHC 2412

England: an automated system is not an agent, as "only a person with a mind can be an agent at law"

– *Software Solutions Partners Ltd, R (on the application of) v HM Customs & Excise* [2007] EWHC 971, at paragraph 67

USA: "a robot cannot be sued"

– *United States of America v. Athlone Industries, Inc.*, 746 F.2d 977, 979 (3d Cir. 1984), U.S. Court of Appeals for the Third Circuit

Germany: machines and software cannot declare intent for purposes of contracting

– *Federal Supreme Court, Judgment of 16 October 2012 – X ZR 37/12*

Thanks to Norton Rose Fulbright LLP

Civil liability: Analogies for causation by machines

USA: Cases relating to Auto-pilots in aircraft

- Claims against manufacturers or operators of planes with auto-pilot-enabled equipment
- Many claims have failed for lack of evidence of manufacturing defects or *lack of proof of causation*

England: law relating to escaping pets/animals

- Animals are, like other chattels, merely agents and instruments of damage, but they are also animate and *automotive*
- An owner of an animal – not the breeder who sells it to the owner – has legal responsibility for the actions of the animal

Germany, US and England:

- Some authorities suggest that, even though a contract may have been entered into *automatically* by software on behalf of a party, it might still be binding on that party.

Thanks to Norton Rose Fulbright LLP

Judicial views of computer decision-making

"A mind of its own?": Some courts are beginning to draw distinction between *deterministic* computers and *AI*:

- **Deterministic systems:** Systems that may be automated but are not autonomous – knowledge assessed at the *time of programming* and by reference to the programmer: *B2C2 Ltd v Quoine Pte Ltd* [2019] SGHC(I) 3 (Singapore International Commercial Court)

- **Autonomous systems**: Would a court look to the opaque subroutines of the algorithm during *subsequent system operation* to determine knowledge?

- **Probabilistic computing:** Computing that is neither deterministic nor autonomous, but based on a *probability* that something is the correct answer. Quantum computing is an example. How would a court deal with *probability* outcomes?

Thanks to Norton Rose Fulbright LLP

Explanations for decision trade-offs

- Imagine being a car driver and facing a difficult trade-off:
  - Stay in left lane, and likely kill a pedestrian
  - Move to right lane, and smash into an oncoming car

- You decide in the moment and end up in court

- You explain your decision, as best you can
  - You decided in the spur **of the moment**

- The court may go down one level of explanation
  - They may examine your state of mind **at that moment**
  - Were you drunk? High on drugs?
  - Were you angry or stressed?
  - Were you insane?

Source: Wikipedia

A self-driving car in court

- Same situation, but now the car was an autonomous vehicle

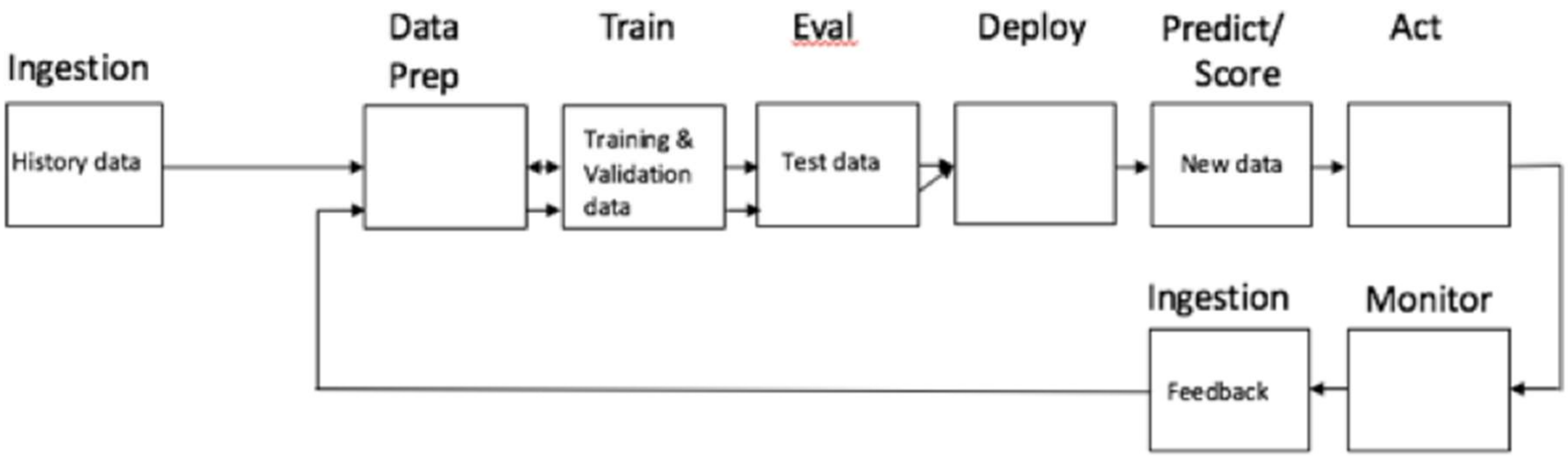
Photo credit: Google 2015

- The court will not accept a statement that the car made the decision in the spur of the moment
  - Because the s/w developers had time to decide what to do in this situation

- The court will examine several layers down to find who or what was responsible, eg:
  - How did the car-control program decide what to do?
  - How did the s/w developers decide how to program the control software?
  - What ethical principles did the s/w developers adhere to (explicit or implicit)?
  - What ethical training had the s/w developers been given?
  - What ethical policies had the car manufacturer or the company employing the developers had in place?
  - Etc.

In practice . . .

Typical data-driven machine learning process



Source: IBM

- Bias can arise with history data, training data, test data, and new data
- Bias can be inserted by the learning process
- Bias can be inserted by the monitoring & feedback activities.

Examples of bias in AI systems

- AI tool for recruitment of software developers at Amazon
  - Amazon gave up attempt after 3 years (2014-2017)
  - Reported by Reuters 11 October 2018

Amazon scraps secret AI recruiting tool that showed bias against women | Reuters

- Automated Bank Loans in US Bank
  - No loans given to people wearing head covering
  - Hats may be a proxy for religious beliefs.

Source: Wikipedia

Recall from Lecture 1: Data-driven vs. Model-driven AI

- Data-driven approaches vs. model-driven approaches

- Machine Learning/ Deep Learning are usually data-driven
  - Patterns are found with no explanation as to why or what these mean

- In model-driven approaches, the AI system has a model of the application domain
  - For example, a causal model connecting causes with effects.
  - Since Windows95, every version of Windows OS has a Bayesian Belief Network linking causes with effects in printer operations, to help diagnose the causes of printer problems.

Identifying bias is difficult in data-driven systems

- We don’t know what factors were used to make the decisions or recommendations

- If the program undergoes evolution or learning, then the developers may not know what code results.
  - Are the s/w developers **responsible** for the code in this case?

- Since we cannot control the output, we focus on what we can control – the production process
  - Looking for bias in the input, training and test data
  - Testing the algorithm for correctness (if we can)
  - Looking at flows of data BETWEEN different AI systems
  - Ensuring good AI Governance

- What comprises good governance for AI systems?

Transparency

- In model-driven approaches, it is usually straightforward to see how a conclusion was reached by the AI
  - We can follow through the IF-THEN rules or reason over the causal model

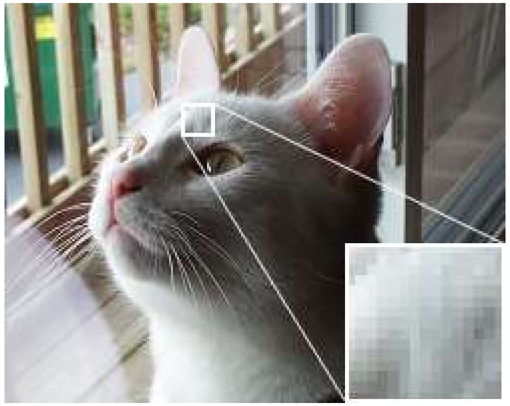
- In contrast, many data-driven approaches are dark (“black boxes”)
  - We cannot see how a conclusion was reached.

- To gain transparency, we may have to build a second AI to mimic the workings of the first
  - A model-driven AI to mimic the workings of the data-driven AI.

Explainability in Model-driven AI – usually straightforward

- Model-driven or symbolic approaches to AI are usually able to generate explanations
  - Because they have a model of the application domain & are transparent
- For example: An Expert System comprising IF . . . THEN . . . Rules
  - IF the patient has lost sense of smell THEN the patient could have CV19
  - IF the patient has a new persistent cough THEN the patient could have CV19
  - . . .
  - IF the patient has all the above symptoms THEN the patient does have CV19
- We can create an explanation for a particular automated diagnosis from the particular IF-THEN rules invoked in the trace of that decision
- Similarly, for other model-driven AI, such as Bayesian Belief Networks.

Explainability in data-driven AI – usually difficult



- In AI methods that are data-driven, such as Neural Networks & Deep Learning methods, the machine is manipulating data without it knowing what the data means

- For example, an image classification program may identify faces by:
  - Examining pixel colours in the image
  - Using pixel colours to identify edges (eg, boundaries of the face)
  - Linking edges together to identify shapes of parts in the image
  - Comparing shapes in image to a library of shapes (eg, chins, ears, eyes)
  - Creating composites of shapes to form faces
  - Comparing faces in different images to find matching faces

- At no point, does the program have any understanding of what is a chin, or an ear, or a face.

- Very difficult to create an explanation for how the decision was reached
  - People don't understand this description of the process.

In addition

- Current Machine Learning and Deep Learning methods are still very immature
  - The resulting systems are not robust to small changes in inputs
  - This makes them easy to hack

- The data-driven approaches require lots of data

- For many situations we do not have enough data
  - Particularly for edge cases and rare events

  (eg, maritime collisions).

Autonomous Vehicles

Sequence of development of AVs:

- Autonomous aircraft (centralized control of airspace, data from isolated experiments)
- Then, autonomous road vehicles (data gained by experiments off-road)
- Lastly, autonomous ships (very little data, no centralized control of high seas).

Source: Rolls Royce

AI Governance

- Companies are starting to put in place processes to govern the creation and deployment of AI systems

- Typically, this will involve a special internal AI Governance committee
  - With representatives of different departments (eg, IT, Operations, Legal)
  - In the best case, including 1-2 outsiders (to avoid "group think")
  - To vet potential AI projects and to oversee their deployment

- Modeled on the Pharmaceutical industry, where these committees are standard

- Companies are also adopting company-wide policies for use of AI
  - Example: Vodafone AI Framework

www.vodafone.com/what-we-do/public-policy/policy-positions/artificial-intelligence-framework

Singapore Model AI Governance Framework

- On 21 January 2020 the Singapore Personal Data Protection Commission (**PDPC**) released the second edition of the Model AI Governance Framework.
- The framework is a voluntary set of compliance and ethical principles and governance considerations and recommendations that can be adopted by organisations when deploying AI technologies at scale. It is not legally binding.
- The Model Framework is based on two high-level guiding principles:
  - Organisations using AI in decision-making should ensure that the decision-making process is explainable, transparent and fair;

  and
  - AI solutions should be human-centric.
- The 2020 edition of the Framework includes real-life industry case studies demonstrating effective implementation of the AI Framework by organisations.

Humans in the loop

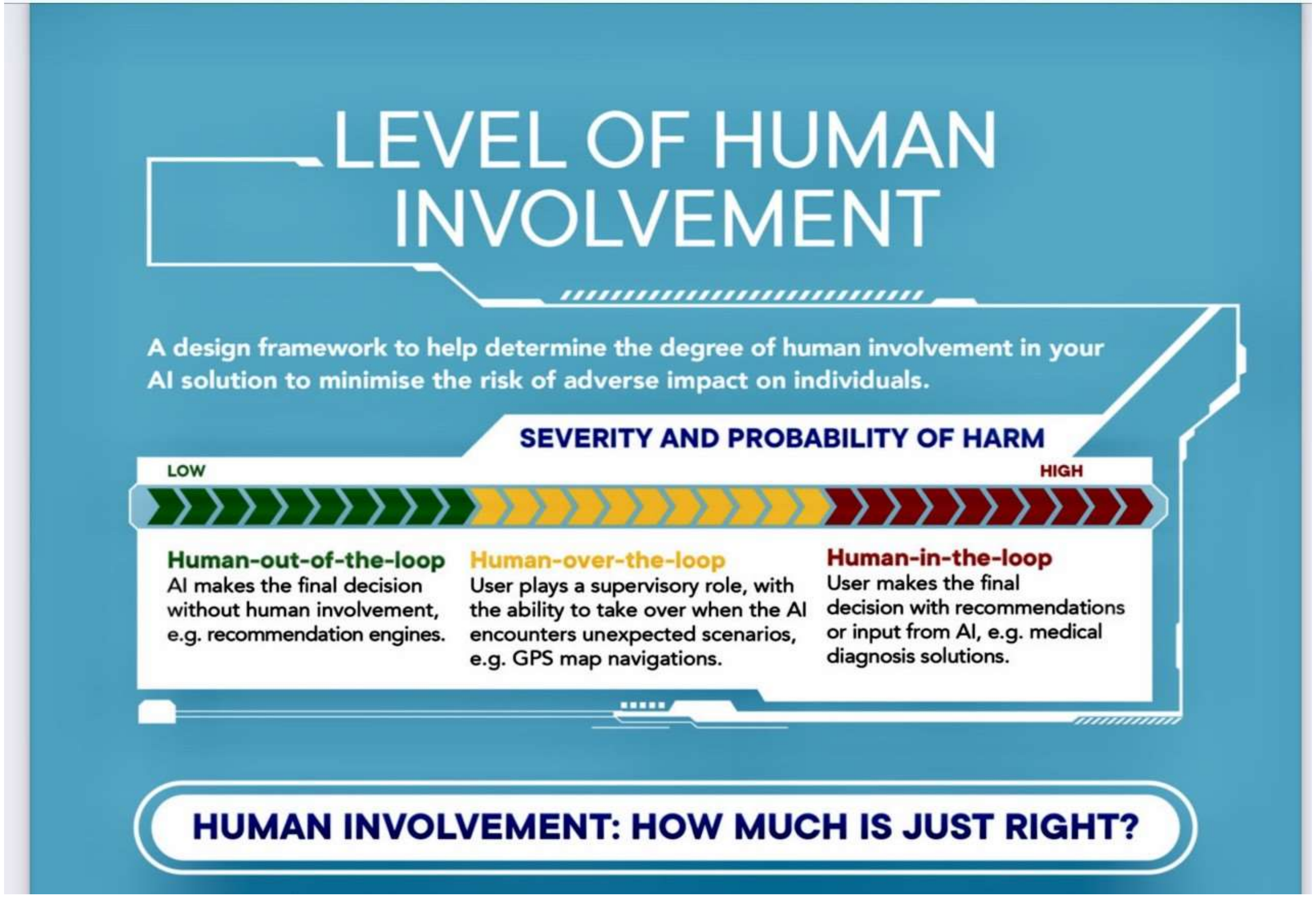
- A key question is to what extent humans should be involved in automated decision-making processes
  - Eg, London Underground – self-driving tube trains on 4 lines, but still have driver sitting in front

- Some regulations only apply to decision-making systems with no humans in the loop
  - Eg, MiFID 2 regulations.

- The human role needs to be sincere (not just for show), or it is likely to be rejected by courts.

- The next slide is a diagram presented in the Singapore Model AI Governance Framework to help companies decide the extent of human involvement in AI decision-making processes.

What level of human involvement is appropriate?



Source: Singapore PDPC
Model AI Governance Framework:
Compendium of Use Cases, 2020.

Some ethical questions

- The tutorials for this week will include some ethical questions
  - There are usually some answers that are definitely wrong
  - There may be more than one answer that is right
  - There may be some answers which are "grey"

- But what is right or wrong?

- Just following orders without question is never right
  - This defence was not accepted in the War Crimes Tribunals after WW II in Nuremberg in November 1945 and in Tokyo in April 1946

- Some situations may require obtaining legal advice

- Many situations can be clarified by discussion (with bosses, colleagues, independent persons).

Reconsidering your orders

"When faced with untenable alternatives you should consider your imperative.“

– Admiral Helena Cain, Battlestar Galactica

Galactica-type Battlestar
Source: galactica.fandom.com

AI and Ethics Summary

- This week we have talked about AI and Ethics
- Why is this topic important
  - Trolley problems
- Regulatory and legal constraints
- Ethical dimensions

  Fairness (non-bias), transparency, explanation, rectification
- Features of this domain
  - ML sensitive, easy to hack
  - Expert systems vs deep learning
  - Humans in the loop
- Deciding what we ourselves should do in specific situations.

Thankyou!