Open-textured regulatory obligations for AI

First thoughts on fairness

Chris Reed

My project

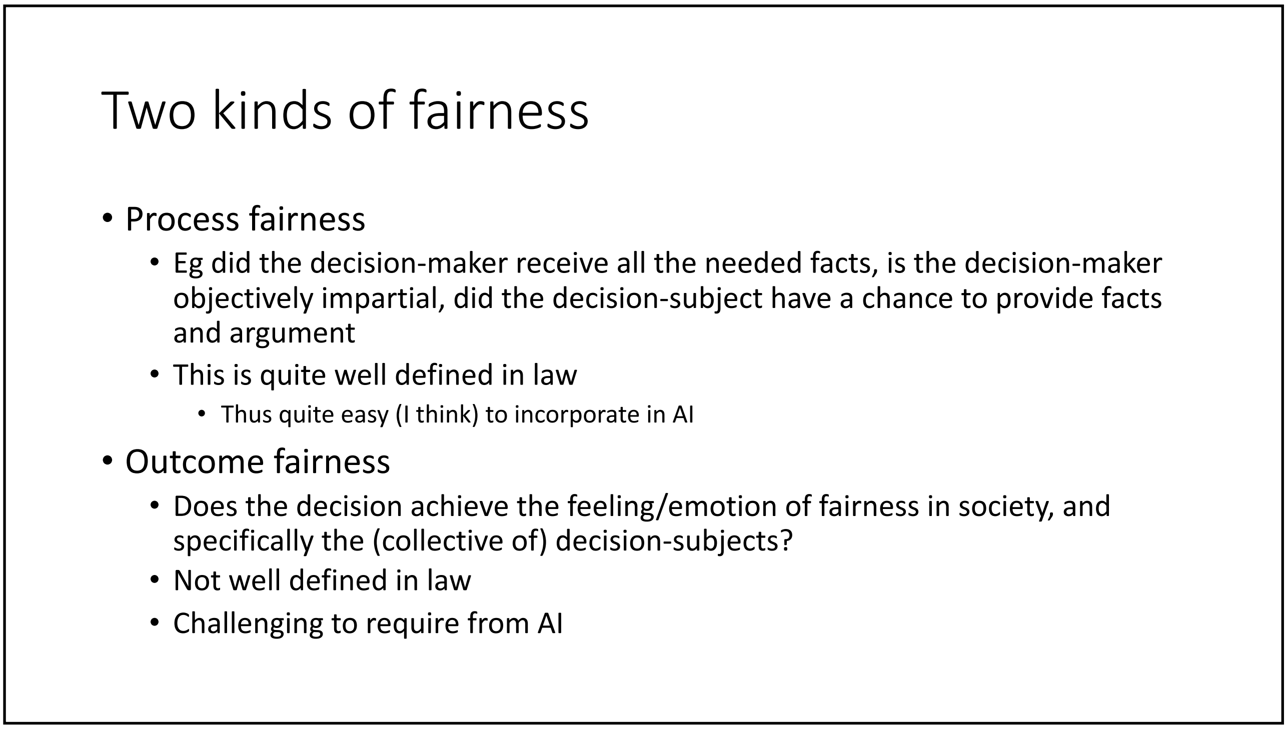
- Assumption
    - Lawmakers will want AIs which make decisions to match or exceed the performance of human decision-makers
        - Human decision-makers are subject to open-textured obligations, to act fairly and reasonably, to achieve safety, etc
    - So we will want AIs to achieve those standard as well
- The problem
    - We don’t define these open-textured obligations in law and regulation
        - We explain them through examples, which are always contextual
        - We assess compliance after the event, rather than defining it in advance
        - We clarify obligations through judicial/regulator decisions, so meaning emerges over time
    - Without clarity of definition, computer scientists can’t produce AIs which meet the open-textured standards

What do I hope to achieve?

- To understand these concepts in greater depth
  - These obligations are imposes to satisfy society’s needs
  - Legal writings focus on the edge cases only
    - So they assume the core content of the concept
    - I need to understand the core as well
  - This means I have to read philosophy, sociology, psychology and economics on these matters to understand the core
- To understand the role which context plays
- If successful, I will be able to explain
  - How law and regulation can be devised in a way which computer scientists can implement in AI
  - The trade-offs which will be necessary
    - Eg when fairness/reasonableness will need to be taken on trust
    - We might need to be satisfied with a minority of unfair/unreasonable decisions

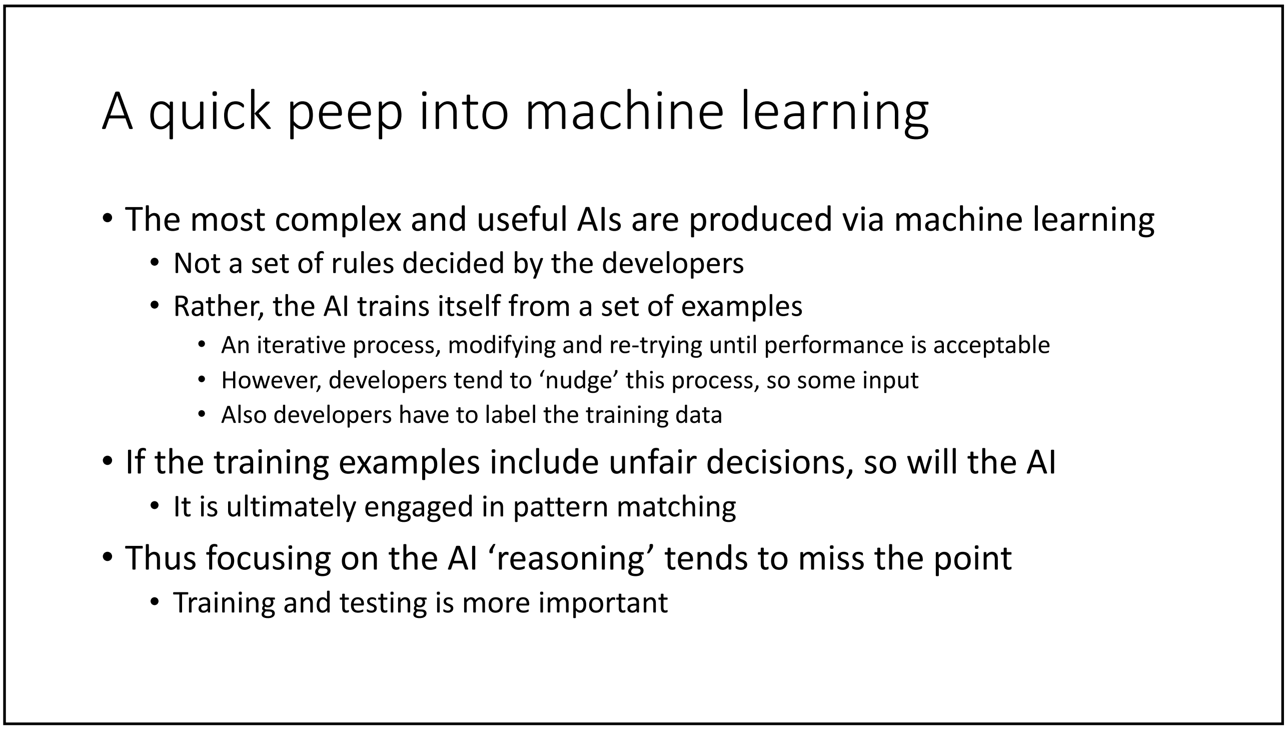
Why start with fairness?

- Possibly the most difficult open-textured concept
  - If I can make sense of it, the others might be easier!
- Computer scientists have researched building fair AIs extensively
  - Thus I can discover where they fail to match the societal and legal conceptions
- Social scientists/philosophers have developed strong understandings of fairness
  - And the role of context in determining fairness
- If I can’t identify how to close the conceptual gaps then this will enable me to move to focusing on alternative mechanisms for fairness
  - Eg human reviews of AI decisions, subject to fairness obligations
    - But human review is expensive and slow
    - It would be better if AIs made fair decisions to start with



Two kinds of fairness

- Process fairness
  - Eg did the decision-maker receive all the needed facts, is the decision-maker objectively impartial, did the decision-subject have a chance to provide facts and argument
  - This is quite well defined in law
    - Thus quite easy (I think) to incorporate in AI
- Outcome fairness
  - Does the decision achieve the feeling/emotion of fairness in society, and specifically the (collective of) decision-subjects?
  - Not well defined in law
  - Challenging to require from AI



A quick peep into machine learning

- The most complex and useful AIs are produced via machine learning
  - Not a set of rules decided by the developers
  - Rather, the AI trains itself from a set of examples
    - An iterative process, modifying and re-trying until performance is acceptable
    - However, developers tend to ‘nudge’ this process, so some input
    - Also developers have to label the training data
- If the training examples include unfair decisions, so will the AI
  - It is ultimately engaged in pattern matching
- Thus focusing on the AI ‘reasoning’ tends to miss the point
  - Training and testing is more important

How do computer scientists deal with outcome fairness

- (Note: this is a simplistic view – I have more research to do here!)
- Fairness is about mathematical ratios between groups
    - Does an aggregate analysis of decisions indicate that one group is advantaged over another?
- Example: a scoring system for job application CVs, those who meet a minimum mark are interviewed

	White	Non-white
Interviewed	18%	16%

	Men	Women
Interviewed	24%	10%

- Discussion

More mathematical puzzles

- Unarticulated assumptions
    - Factor scores correlate with ability/fitness (whatever is decided as the basis for the AI)
    - Choice of the appropriate factors to control for is objective
    - Equal ratios = fairness to individuals
        - Collective v individual fairness has received detailed legal attention
        - Computer scientists have developed some ideas on individual fairness, but …
- Deciding between competing fairness norms (eg race and sex equality)
- Implementing societal policies
    - Weighting for disadvantage
        - Is a high score by a less-educated person ‘worth’ more than the same score by a higher-educated person?
    - Affirmative action?
- More to research here
- But it should be obvious that ratios between groups is not (completely) how societal fairness, or legal fairness, works

Outcome fairness in theory and in society

- What have I learnt from social scientists and philosophers?
- Fairness is about sharing of resources
  - Money
  - Employment
  - Opportunities (eg medical treatment, education)
  - Etc
- Voluntary sharing v sharing decision-makers
- Three main aspects to be considered
  - Equality of treatment
  - Power and status
  - Making fairness convincing

Equality of treatment

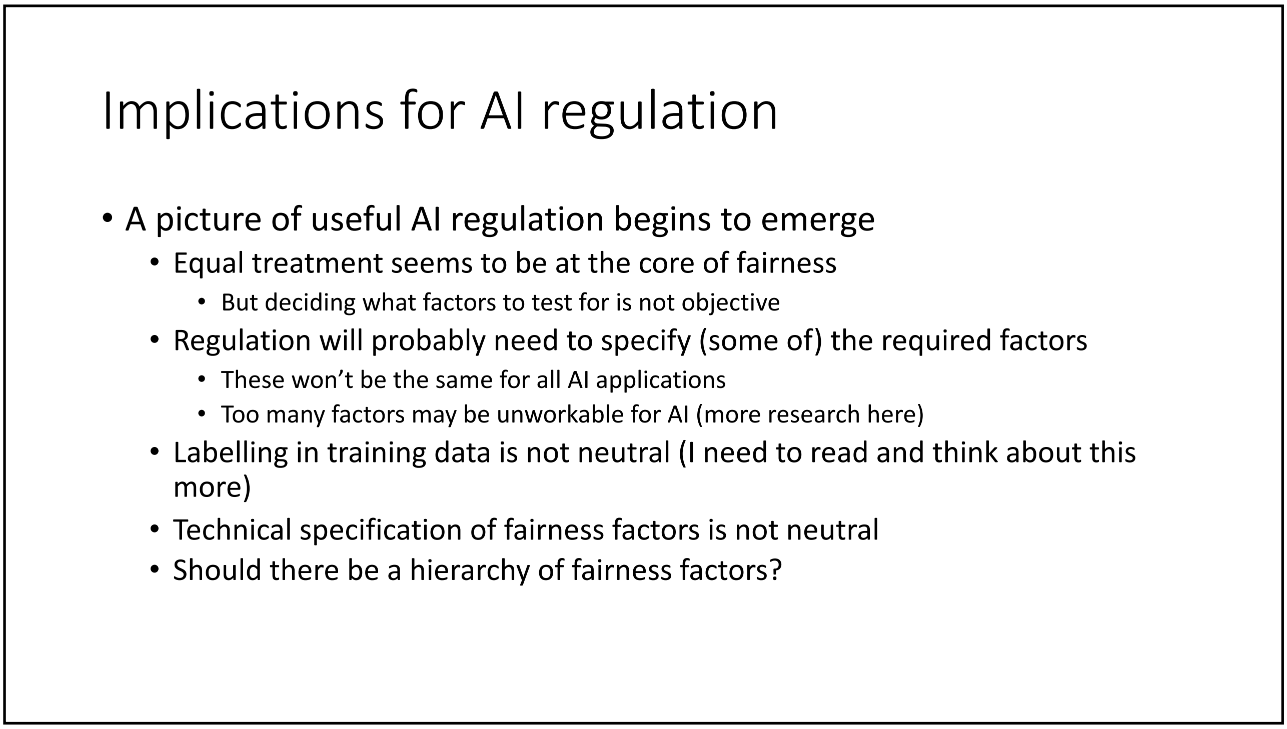
- Fair sharing requires treating individuals like other members of their sub-group (equality) – BUT
- Individuals may deserve different treatment in some cases, depending on merit
- Merit is not an objective state
  - What amounts to merit is culturally/socially determined. All these are seen in some societies as deserving a greater share:
    - Innate ability (strength, skill, cleverness, etc)
    - Effort
    - Choices made
    - Social status and power
    - Disadvantage
  - What amounts to merit may vary with context (eg in good times effort deserves more than poverty status, in a famine poverty status outweighs effort)
- Fair sharing also requires sharing in accordance with merit as between sub-groups
- The criteria for disadvantages which a decision maker needs to control for are in part socially determined
  - How far should law counter this?
  - Machine learning issue - legal norms like fundamental rights are necessary because humans don’t sufficiently follow fairness norms in practice

Power and status

- Sharing in operation is influenced both by power and by status.
  - Power tends to gain a higher share
    - Whether this is fair differs between cultures
  - Status also gains a higher share
    - This seems to be considered fair by those who expect to share
  - Do legal norms reflect this, or should they correct for it? This might also be a cultural question
- When allocation is undertaken by a decision-maker (rather than collective social norms) that decision-maker has power over the rest
  - Power to determine the merit elements of the fairness criteria
  - Power to to determine the process of decision making
- Is one role of the law to control exercise of power?

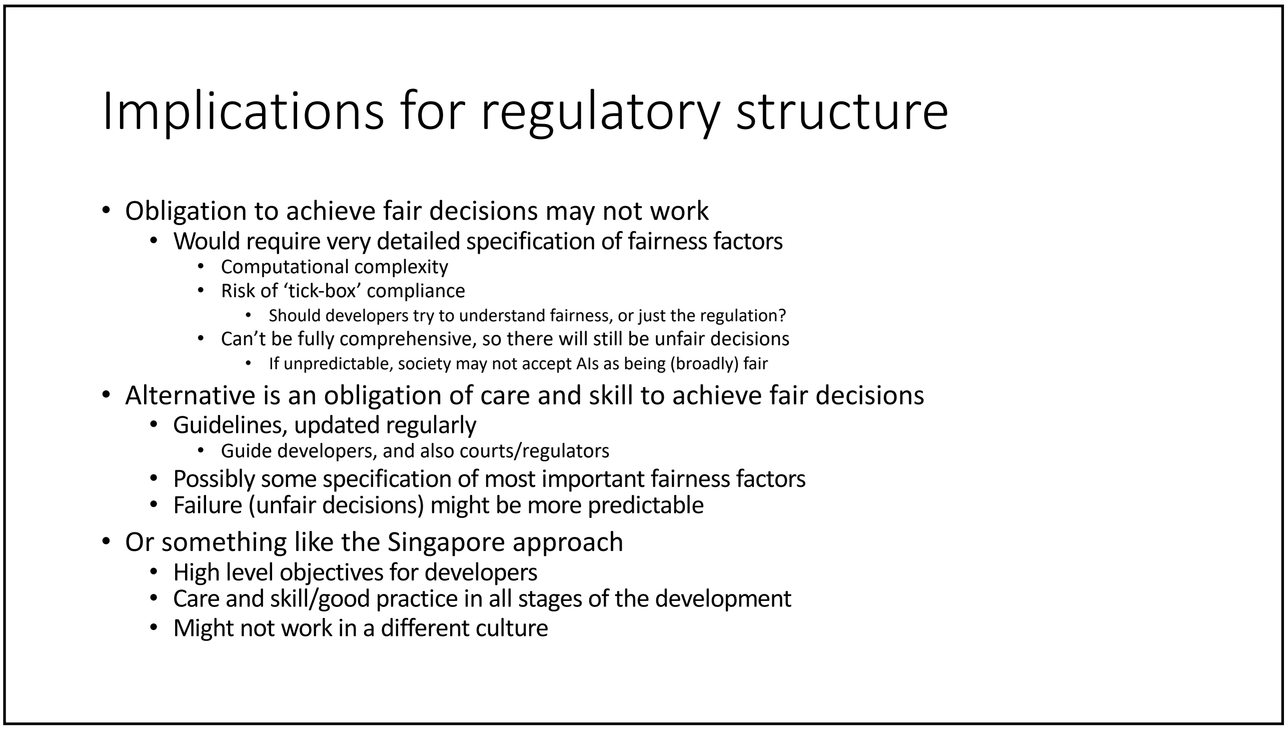
Making fairness convincing

- Judgments about fairness are not absolute, but instead they are comparative
  - Comparison requires information about how others have been treated
    - In terms of outcome
    - In terms of procedure
- All fairness theories propose that fairness is not an objective concept
  - Fairness is a subjective state or quality in the minds of subjects of decisions
    - Based on both the decision outcome *and* the decision process
  - Is it enough for the majority of subjects to accept that the decision was fair?
    - How does this relate to legal tests for fairness?
  - Decision subjects can see the process, but may only see a single decision
    - Thus procedure may play more weight in their assessment
    - Subjects may believe that a fair procedure leads to fair outcomes if they can't see a range of outcome decisions
- All fairness norms are culturally determined, so this might affect assessment of fair process as well



Implications for AI regulation

- A picture of useful AI regulation begins to emerge
  - Equal treatment seems to be at the core of fairness
    - But deciding what factors to test for is not objective
  - Regulation will probably need to specify (some of) the required factors
    - These won't be the same for all AI applications
    - Too many factors may be unworkable for AI (more research here)
  - Labelling in training data is not neutral (I need to read and think about this more)
  - Technical specification of fairness factors is not neutral
  - Should there be a hierarchy of fairness factors?



Implications for regulatory structure

- Obligation to achieve fair decisions may not work
  - Would require very detailed specification of fairness factors
    - Computational complexity
    - Risk of 'tick-box' compliance
      - Should developers try to understand fairness, or just the regulation?
    - Can't be fully comprehensive, so there will still be unfair decisions
      - If unpredictable, society may not accept AIs as being (broadly) fair
- Alternative is an obligation of care and skill to achieve fair decisions
  - Guidelines, updated regularly
    - Guide developers, and also courts/regulators
  - Possibly some specification of most important fairness factors
  - Failure (unfair decisions) might be more predictable
- Or something like the Singapore approach
  - High level objectives for developers
  - Care and skill/good practice in all stages of the development
  - Might not work in a different culture

Where next for the research?

- A starting assumption was that AI decisions need to align with societal fairness norms
  - Obviously not fully achievable
    - Norms are too diverse, and depend on both context and culture
  - But some mismatch might be acceptable
    - Mandatory imprisonment laws are unfair to a minority (individual) but perceived as fair for society as a whole (group)
  - Law has focused on a subset of fairness norms
    - Strong rules on procedure
    - Is it clear about outcomes? I need to find out
- AI development has focused on group fairness norms
  - How far can it capture context and culture?
  - Can it incorporate individual fairness?
- What is context?
  - Might it be captured by the choice of fairness factors?
- Once these are answered, it might be possible to identify how law can mandate fairness in AI decision-making
  - And the limits of what law can ask AI to achieve

Questions, suggestions and ideas