ASSUMPTIONS OF MAKING THINGS COMPUTABLE

Mireille Hildebrandt
IMPLICATIONS OF MAKING THINGS COMPUTABLE

Mireille Hildebrandt
My 3 AI cards on the table

1. Things that matter are not computable
2. They can nevertheless be made computable
3. They can be computed in different ways and that difference matters
Does the political economy of RL concern the scientific question of:

1. determining the limits of the reward hypothesis for a given domain?

- reward hypothesis: all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward).

Or does the political economy of RL concern the political questions of

1. who get to decide the difference that makes a difference?
2. how the proxies for specified goals are chosen and defined?
3. where QO fits in, and what this means for democracy?
Is the political economy of RL built on:

- Utilitarianism, behaviourism and law & economics?
  - As in rational choice theory and/or behavioural economics
  - That inverse the relationship between proxy and what it stands for
Or, is the political economy of RL built on:

- Polanyi’s *The Great Transformation and further development of his analysis*?
  - Highlighting the nefarious implications of the hegemony of economic markets
  - Rearticulating the economy as embedded in society, subordinating economics to the political, the social and the ethical
  - Turning the reward hypothesis inside out as to what ‘proxies’ for what

The Issue of Proxies, And why EU law matters for recommender systems

AUTHORS
Mireille Hildebrandt

Abstract

Recommendations are meant to increase sales or ad revenue, as these are the first priority of those who pay for them. As recommender systems match their recommendations with inferred preferences, we should not be surprised if the algorithm optimises for lucrative preferences and thus co-produces the preferences they mine. This relates to the ...

See more

Paper DOI
The issue of proxies

Both behaviourism and the reward hypothesis depend on:

- *Inversion* of the relationship between a concept/practice/institutional fact
- And the mathematizable proxy for that concept/practice/institutional fact
- *Deciding on the proxy* (labelling in SL, training data in USL, goal in RL)
- Involves political choices, that require deliberation and participation
4 The Issue of Bias: The Framing Powers of Machine Learning

Mireille Hildebrandt

4.1 Productive Bias, Wrongful Bias, and Unlawful Bias

In this chapter I will discuss three types of bias and their interrelationship. The first concerns the bias that is inherent in machine learning. This type of inductive bias is inevitable and, though neither good nor bad in itself, is never neutral in real world settings. The second concerns the bias that is problematic from an ethical perspective because it (re)configures the distribution of goods, services, risks, and opportunities or even access to information in ways that are morally problematic. This may regard categorical exclusion of people or the softer tyranny of nudging people into a certain direction based on traits or behaviors. Let’s note that these traits or behaviors may be observed (by sensor technologies or online tracking systems) or inferred (by way of machine learning). Bias in observation affects the training data, and bias in inferences affects the throughput of the system; both impact the output. The third type of bias concerns unlawful bias, that is, the targeting of people based on prohibited grounds. This may be a subset of ethical bias, but sometimes bias that is not ethically problematic may nevertheless be unlawful because, for instance, discrimination on the basis of gender may be prohibited categorically, even if some would argue that there is no ethical implication (e.g., charging men a higher car insurance premium because they are found to be more risk prone than women is not necessarily an ethical problem).
The issue of bias

- Wolpert’s NFL theorem
- Gadamer’s ‘prejudice’
- Popper’s theory-laden perception
- Hume’s scepticism

- Bias is inherent in living agents:
- They need to detect the difference that makes a difference, but …

Reward functions

- Who get to decide on the goal?
- Who decide on how e.g. a state relates to the goal?
- Which are the assumptions of ReLe?
- How do ReLe assumptions relate to ReLi?
- Who says?
- Who decides?
Law as a reward function

- What if the political economy of law demonstrates that
  - Corporations are rewarded if they take us for a ride - hook, line and sinker?

- What if we get to rewrite the law such that
  - Corporations are rewarded for doing the right thing?

- Who is ‘we’?
- Who get to decide what is the right thing?
- Do ‘we’ know what is the right thing in case of things that matter
  - (for whom, in the long run, at what cost, for whom, in the long run)?
The EU proposal for an AI Act

- Imposing stringent quality control, risk assessment, data governance, robustness, accuracy, cybersecurity and human oversight on providers of high risk AI systems

- Transforming the economic incentive structure for the development, provision and deployment of high risk AI systems

- Reconfiguring the political economy of RLSs
The EU proposal for an AI Act

In the recently released ‘compromise’ draft of the Council of the EU:

- Provision of general purpose AI is excluded from the scope of the Act
- Making those who depend GPAI for the development of AI systems
- Responsible for the design decisions
  - taken by big tech hegemons capable of developing GPAI
The rule of law (‘Rechtsstaat’) concerns the institution of countervailing powers
- *Who get decide what when and how*
- *Note the link with Polanyi’s countermovements (not the same but…)*

The rule of law
- *Does not decide the goal but decides the space for the reward function*
- *Thus also constraining the space of the goals that can be determined*
The determination of goals and reward functions in RL systems that impact RL

- Should be subject to participation by those who will suffer the consequences
- By those who are ‘made computable’
- We must learn – as a society, as a polity – to decide on how to make the design decisions that matter
- In ways that reward diversity and inclusion
The End