SafeAI22 (AAAI-22)

# The problem of behaviour and preference manipulation in AI systems

Hal Ashton, Matija Franklin

Causal Cognition Lab, UCL
Department of Computer Science, UCL

# Genesis of Paper

Stuart Russell, Human Compatible: AI and the problem of control (2020):

(On the subject of Recommender Systems):

*”Like any rational entity, the Algorithm learns how to modify the state of its environment - in this case the user’s mind - in order to maximise its own reward”.*

This is an example of the Observer effect - The ML system is often not neutral to the preferences that it learns.

# Structure

- The treatment of preferences in ML applications
- The relationship between Preferences and Behaviour
- Mechanisms that change manipulate preferences
- Application to Value alignment
- New Research directions

# Preferences in ML

# Preference Learning in ML

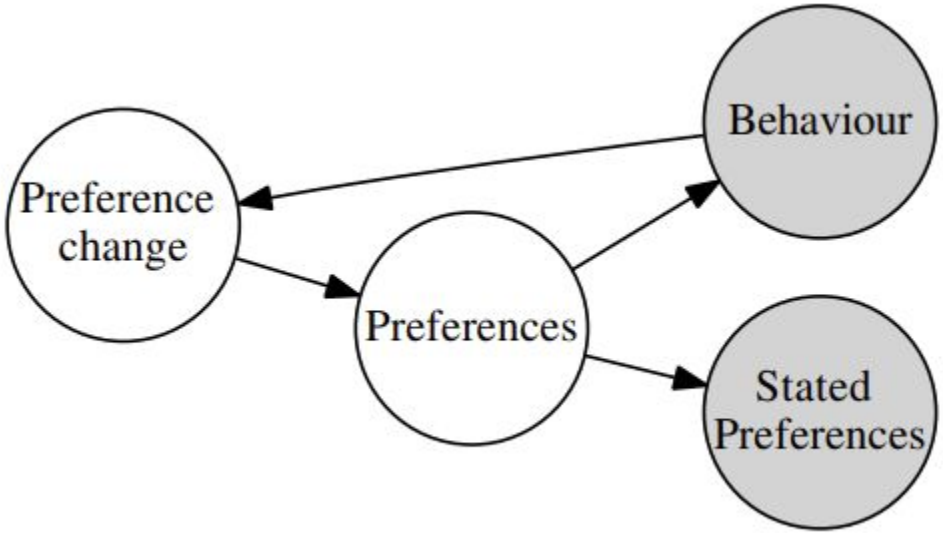
- ➔ Machine Learning is often used to learn the preferences of users in order to better deliver some service to them.
- ➔ Preferences can be learned directly by asking subjects (Stated Preferences) or inferred from behavior (Revealed Preference).
- ➔ Both have limitations (e.g., 'irrational' behavior, environmental effects).
- ➔ **Preferences are not static**

Behaviour change accepted; preference change unacknowledged

# Preference Change and Behaviour Change

➔ The behavior change practices cause a preference change as well

➔ Behavior history forms preferences



# Behavior Change

- ➔ Behaviour can be reliably changed with a variety of well researched techniques.
- ➔ Behavior change complex:
  - ◆ Advertising Industry (commercial)
  - ◆ Behavioral Science (academic)
- ➔ Popularised by nudging - changing choice architecture to change behavior without limiting or forcing options.
- ➔ All environments influence behavior, but certain environments are more influential than others.

# Preference and AI

- ➔ Problem - The iterative nature of ML
- ➔ AI/ML systems learn preference and change their interaction with a human in line with those preferences. This change in interaction influences human behavior.
- ➔ As many AI/ML systems influence human behavior, they also influence human preference.
- ➔ By learning preferences, the AI is changing preferences.

# The mechanisms that manipulate preference

# Popularity Bias in Recommenders

- ➔ Certain popular items are recommended more often than less popular items, this increases their chance of being recommended by more users, allowing them to grow even more popular - the process reinforces itself

- ➔ Symptom of **confounded data**:
  - ◆ Behaviour data used to train and test algorithms has already been influenced by the algorithm
  - ◆ Creates an amplifying feedback loop which increases homogeneity of recommended content
  - ◆ Naive recommenders will push towards the same small set of content

# Psychological Mechanisms

- Biases: mere-exposure, availability, recognition:
  - The more we are shown a certain type of content, the more we accept and like it.
- People can even be told what their preferences are and they will change to match:
  - **Hall, Johansson, and Strandberg 2012**: Even after having given their preferences, when they were secretly changed by the experimenters, participants would often alter their views to match their (falsely recorded) ones
- Content types which engenders strong emotion is more addictive.

# Popularity Bias + Exposure bias = User homogenisation

- By recommending users a narrow selection of content, content recommenders make users more predictable.
- Though often unintended, the changes brought in user preferences are favourable for system owners, principally by making users more predictable.
- The algorithms are doing what they were designed to do - make money efficiently for their owners by increasing the time their users spend online.
- The measurement and valuation of the harm caused is difficult
- Would an AI system be incentivised to intend to change human preferences?

# Value Alignment and Preferences

# Inverse Reinforcement Learning (IRL) and Alignment

- Construction of a human's utility function or values by the observation of their behaviour
- IRL are a putative solution for the problem of value alignment:
- Russell (2000) suggests three principles for AI developers to create beneficial machines which all rely on preferences
    1. The machine's only objective is to maximise the realization of human preferences.
    2. The machine is initially uncertain about what those preferences are.
    3. The ultimate source of information about human preferences is human behavior
- The difficulty in applying these principles is the causal relationship between behaviour and preferences

# Summary and Research Directions

# Problem Summary

1. Users will interact an adaptive system over time and their preferences and behaviour is influenced by that system.

2. It becomes impossible to know whether the system is doing a really good job or whether the system has just altered the preferences of its users to do a really good job.

# Related Research and Contact

Contact us:: Matija.Franklin@ucl.ac.uk

Hidden Incentives for Auto-Induced Distributional Shift (Krueger, Maharaj, and Leike, 2020)

Agent Incentives: A Causal Perspective (Everitt et al, 2021)

User-tampering in Reinforcement Learning Recommender systems: (Evans and Kasirzadeh, 2021)

# Thank you

matija.franklin@ucl.ac.uk