The Moral Psychology Behind Artificial Moral Advisors

Yuxin Liu

Moral Psychology of AI workshop

June 26, 2023
Kent, UK

Centre for Technomoral Futures

THE UNIVERSITY of EDINBURGH
School of Philosophy, Psychology and Language Sciences

Moral AIs

Artificial moral agents (AMAs), moral AIs, moral machines, etc

- Machines endowed with moral reasoning capabilities

For	Against
- Inevitability/necessity	- Machines cannot be moral
- AI is too complex to understand	- No universal agreement in ethics
- To prevent immoral use	- Existential concerns/risks
- To ensure humans' safety	- Slave argument
- Ethical alignment with humans	- Moral deskilling
- Enables moral justification/explanation	- Responsibility/accountability issues
- Increases public trust	- Moral reasoning does not mean ethical behaviours
- Makes better moral decisions	- Undermines human moral agency
- Leads to better understanding of human morality	- Value imperialism
- etc.	- etc.

One major motivation

Machines make better moral decisions than humans do

- Unaffected by human biases and heuristics
- Can be used for AI moral enhancement – improve human's morality through technological means

Ideal observer theory (IOT)

Firth (1952)

1. Omniscient: possessing factual knowledge of all non-morally relevant information involved in the procedure of deciding the rightness/wrongness of a particular act
2. Omnipercipient: capable of simultaneously imagining or visualising all alternatives and consequences of any given act
3. Disinterested: completely impartial about reacting (dis)favourably to any person/thing
4. Dispassionate: incapable of any emotional experience at all
5. Consistent: having exactly similar ethically significant reactions to any given act
6. Normal: in other aspects as a human being

Moral AIs as ideal observers

Giubilini & Savulescu (2018)

- "The AMA would implement a quasi-relativistic version of **the ideal observer** … the AMA is disinterested, dispassionate, and consistent in its judgments."

Sinnott-Armstrong & Skorburg (2021)

- "Such an AI could serve as a proxy for **an ideal observer** or at least evidence of how an ideal observer who is informed, rational, and impartial would make moral judgments and decisions in cases like these."

Survey of machine ethicists (Martinho et al., 2021) – machines makes better moral decisions as they are consistent, dispassionate, and impartial

- Machine objectivism: "Through logic and context-specificity, they are better moral reasoners and educators."

Today's talk

Moral AI as ideal observer for the purpose of helping people achieve moral enhancement?

I. Moral AI & the **I**deal **O**bserver **T**heory
II. Moral AI as ideal observer?
III. Responding to moral AI

1. Omniscience & 2. Omnipercipience

IOT is unattainable

- Humans are epistemically bounded

IOT's circular reasoning

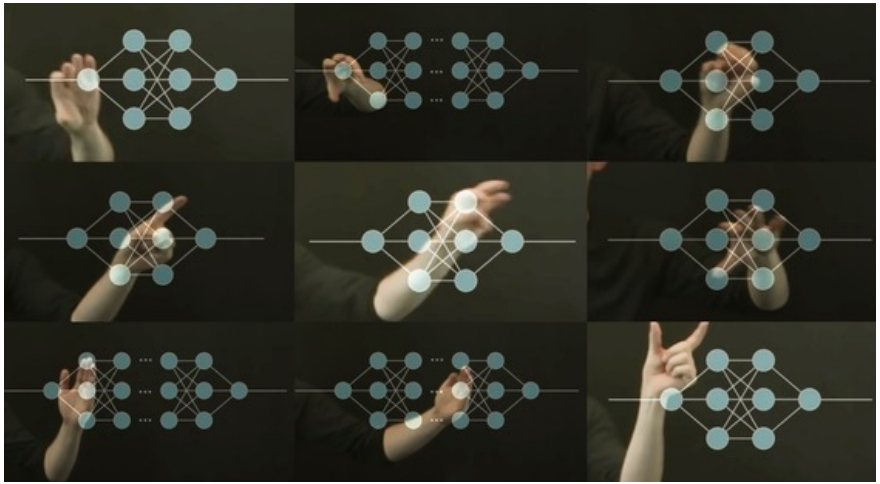
- The view from nowhere (Gebru, 2020)
- The bootstrapping problem (Vallor, 2016)

Human moral fallacies
↓↑
Need for moral machines

3. Consistency

Consistency reframed as (dis)similar features

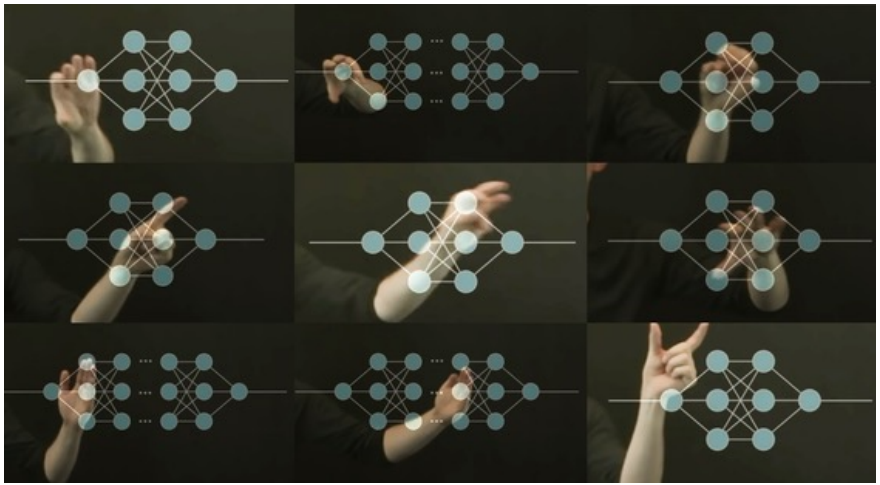
- Two actions, however similar they are, are dissimilar in at least one aspect (Harrison, 1956)
- Some features should/should not be influenced by emotional/ intuitive reasoning



4. Dispassion

Dispassionate machines ≠ superior reasoners

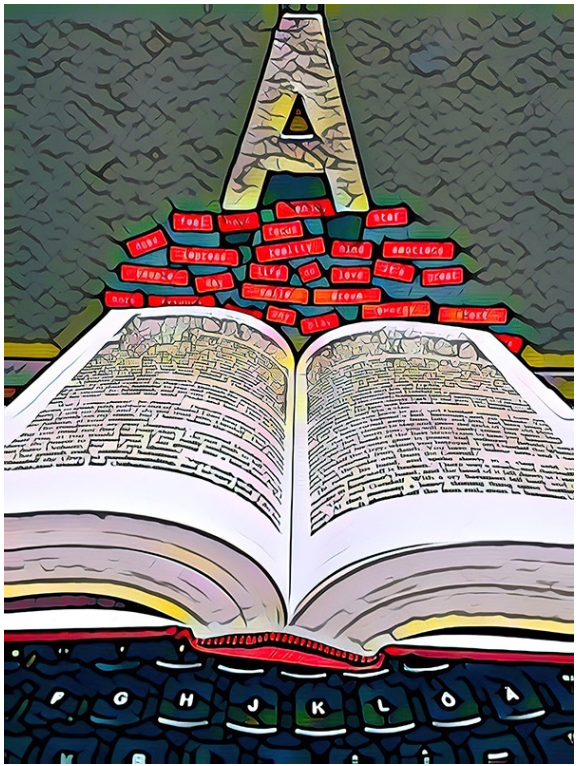
- Bootstrapping again – bias/ prejudice seeping through machines
- Blanket rejection of intuitions creates a hierarchy of values
  - Incompatible with human moral psychology
  - Undesirable due to lack of flexibility



5. Impartiality

Inherent political properties of moral AIs as a product of intentional human creation (Winner, 1980)

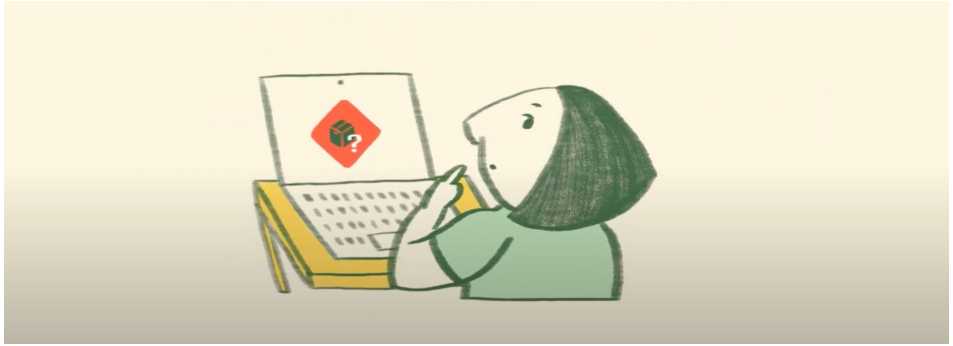
- Political by explicit/implicit design
  Deliberately/inevitable benefit some and disadvantage others
- Political by necessity
  Technosolutionist nature of moral machine projects



Responding to moral AIs

Passive acceptance (Lara & Deckers, 2020)

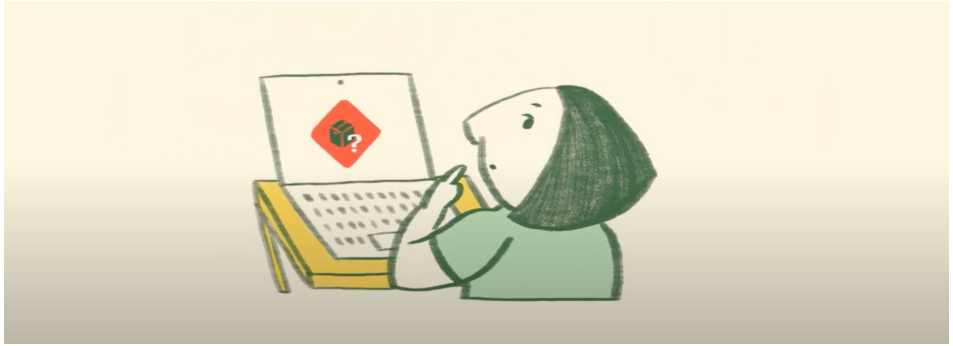
- Genuine moral enhancement?
- Risks of moral deskilling from outsourcing (Vallor, 2015)
- But is it the only option?



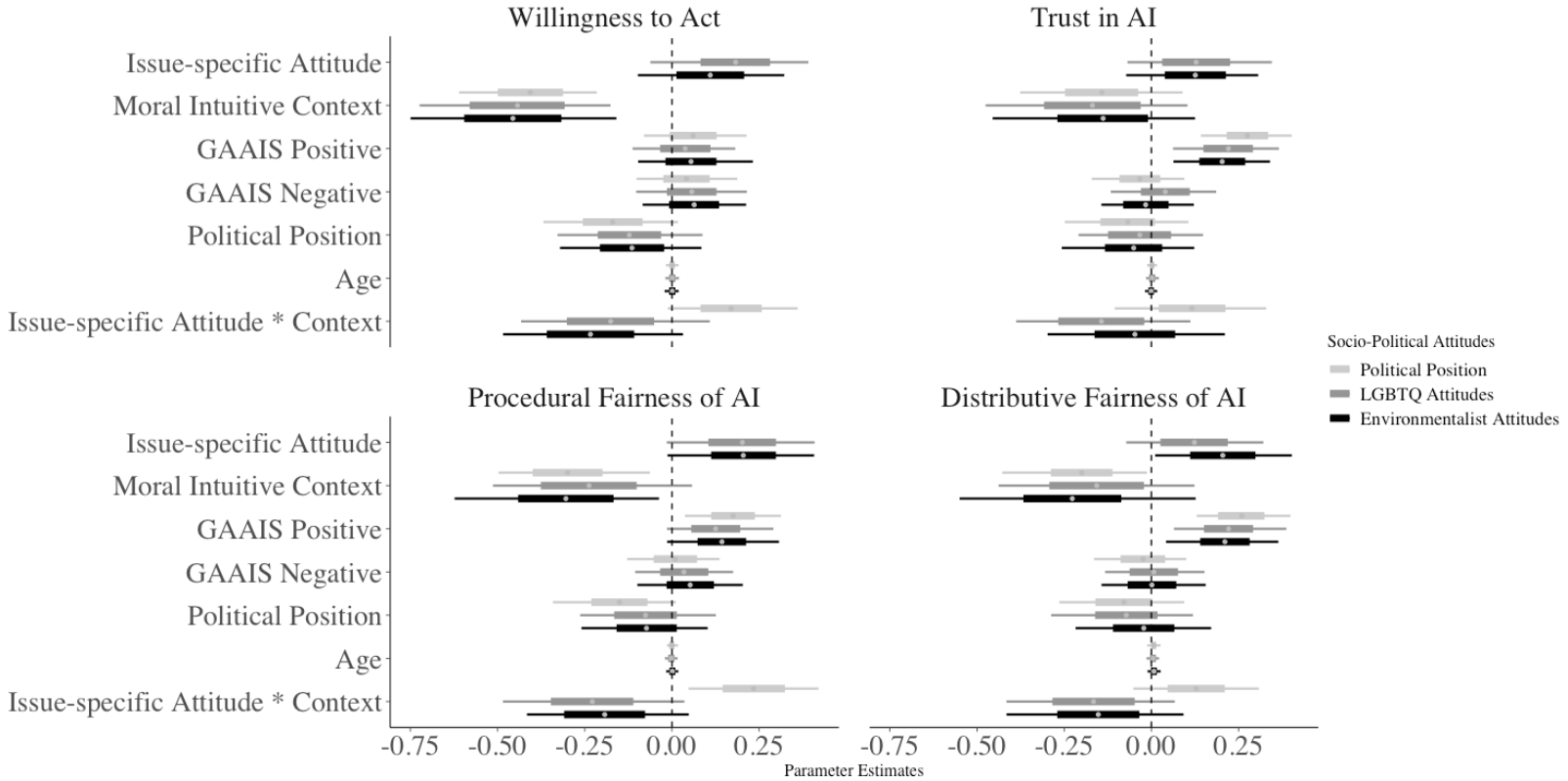
Responding to moral AIs

Passive acceptance is not the only option – inescapability and irreducibility of human decisions

- Sartre's advice-seeking soldier
- Prescriptive nature of AI moral advisors
- Diminishing benefits of moral AIs
- Possibility of moral degradation from motivated reasoning



Moral judgements towards AI



Conclusions

Is all hope lost? AMAs may be useful

- As an information provider
- In specific contexts, instead of a general-purpose moral advisor
- In non-emergency situations

Moral AI to help humans achieve moral enhancement?

- Essential humans input in the design of moral machines
- Inescapability and irreducibility of human moral decisions

Partially based on:

Liu, Y., Moore, A. Webb, J., and Vallor, S. (2022). Artificial moral advisors: A new perspective from moral psychology. *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 436-445. https://doi.org/10.1145/3514094.3534139

AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY


yliu3310@ed.ac.uk


yliu-psych.github.io


@_yuxinl_