On Moral AIs

Yuxin Liu

Society for Philosophy and Technology
23rd Biennial Conference

Panel: Mobilising Technomoral Knowledge

June 9, 2023
Tokyo, Japan

Centre for Technomoral Futures

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

Moral AIs

Artificial morality, moral AIs, moral machines, artificial moral agents, artificial moral advisors, etc.

- Mainly concerned with implicit/explicit ethical agents (Moor, 2006)

One major motivation in machine ethics

- AIs are superior moral reasoners over humans

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

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

Omniscience & 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

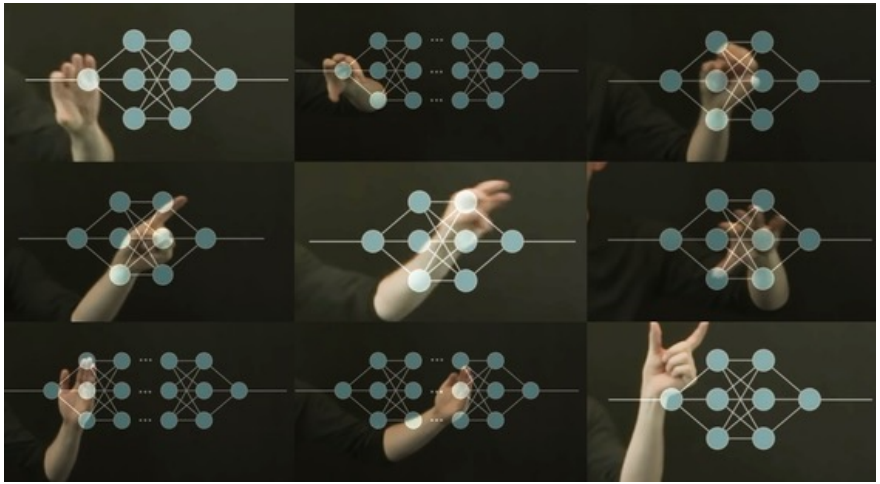
↓↑

Need for moral machines

Consistency

Consistency reframed as (dis)similar features

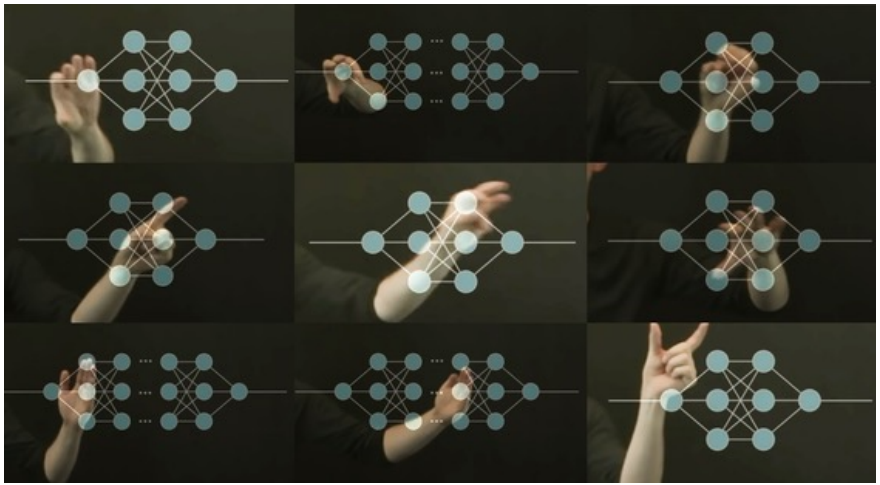
- Two actions, however similar they are, are dissimilar in at least one aspect (Harrison, 1956)



Dispassion

Dispassionate machines ≠ superior reasoners

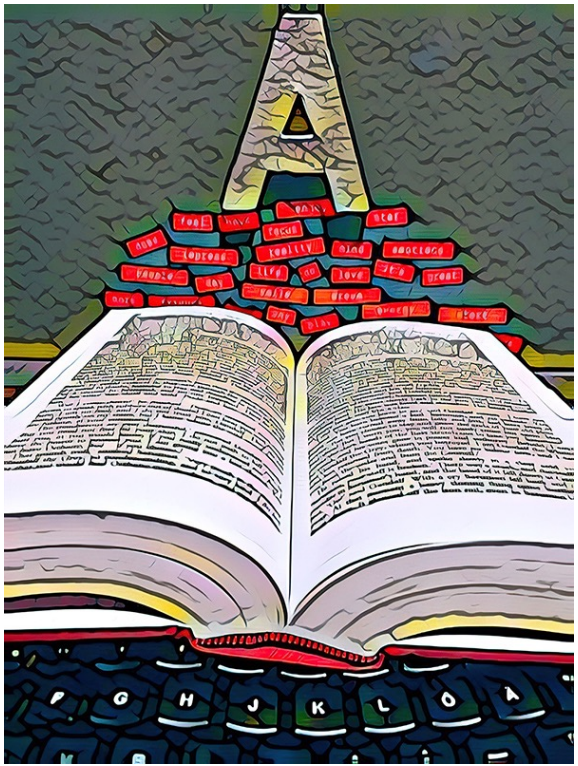
- Bootstrapping again – bias/ prejudice seeping through machines
- Creates/manifests a hierarchy of values – blanket rejection of intuitions



Impartiality

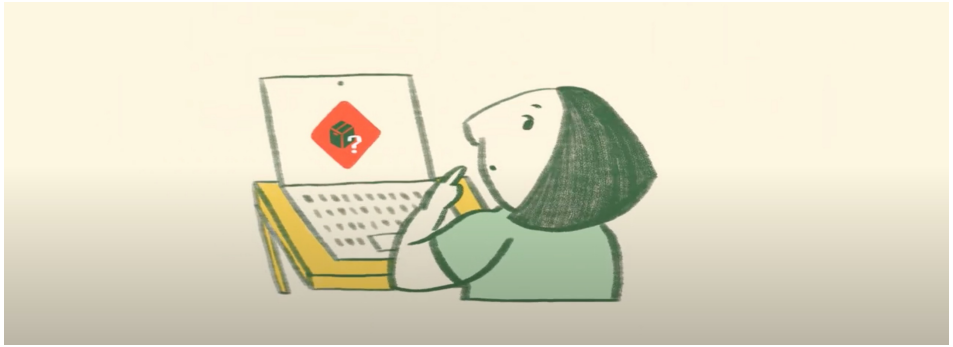
Inherent political properties of moral AIs as a product of intentional human creation (Winner, 1980)

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



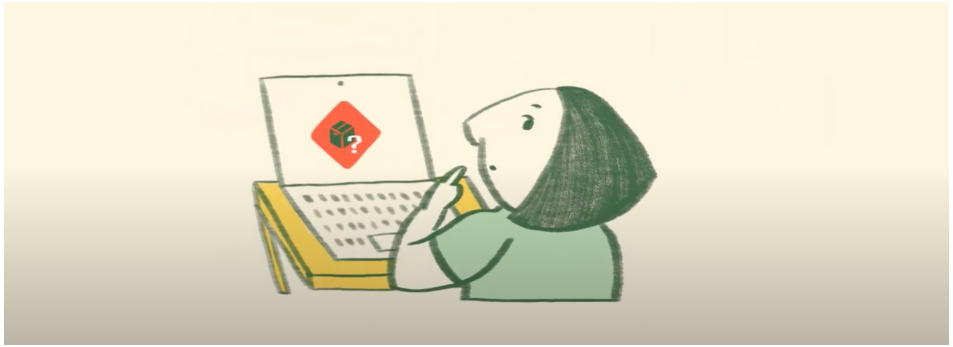
Responding to moral AIs

- Passive acceptance
  - Genuine moral enhancement?
  - Risks of moral deskilling from outsourcing
  - 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



Necessity for technomoral wisdom

Moral AI as an ideal observer?

- Bootstrapping problem
- Misconception of appropriate decision making
- Political properties of moral AIs

Moral AI to help humans achieve moral enhancement?

- Inescapability and irreducibility of human moral decisions

Thank you

Centre for Technomoral Futures

technomoralfutures.uk

yliu3310@ed.ac.uk

@_yuxinl_