AI, Politics and Power

A talk by Carla Zoe Cremer,
Winter Fellow,
GovAI

18:00-19:00, followed by social
Wednesday, 4 May 2022
Large Lecture Theatre - Department of Statistics
(OX1 3LB, Banbury Road entrance)

Seminar series by the OxAI Safety Hub

Intro

level of abstraction guarantees inaccuracies

Another framework that passes the threshold of not being correct but providing useful questions

Solutions to AI risk depend on threat model

Risk

originally both unlikely disaster and fortune

probability distributed over some outcome

negative outcome according to who?

reducing risk = shifting distribution

for and despite other people who don’t get to shift it

if you want to shift risk, you want to exercise or shift power

AI Alignment - Power revisited

power as exercising *control* over the probability of outcomes that are preferred by the controller

AI risk is largely revisiting old problems encountered by co-existing, differing, agents (human-human; human-institutions; human-God?)

from laws to code

technical alignment is a subset of alignment

AI Alignment

a broader view of AI alignment as the alignment of algorithmic decision-procedures with the interests of the polis (a political unit of agents whose preferences we care about)

you cannot ‘solve’ risk. you can redistribute it by exercising / shifting control

economics of control via AI are favourable

-> leverage / ‘impact’

AI as effective control

control via time-keeping, classifications, tracking of locations and preferences, correlating, inferring, predicting, detecting

inputs to control : accuracy + predictability + replicability

who has access to production, goal setting?

what outcomes do they favour?

what barriers do they face on inputs?

Crawford 2021, Chapter 2 ; Ahmned 2021,https://magazine.scienceforthepeople.org/vol24-3-cooperation/we-will-not-be-pacified-by-participation/

AI as politics

Automated scaled decision-making

AI as a generalisable codification of rules that govern choices of many agents (who differ)

expertise/ preferences / ignorance of the few can be imposed on everyone else

---

those who decide on the risk distribution = those affected by the outcome?

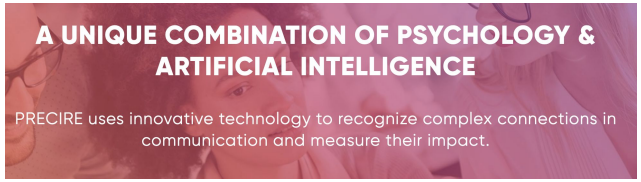
recap

code encourages standardisation, discretisation, simplification, classification and metricization

coded DM could enhance fairness via standardisation and rule transparency

– failures and benefits depend on where that is applied

– what procedures ensures we apply it well?



Thread model 1

expect deployment before completion

simplicity fails to capture the task

goodhart’s law

human-specific classification boundary

short term

worst case: controller ignorance

AI & Accuracy

accuracy (understanding & capacity) as input to greater control

one of the most dominant drivers of research directions, design and use?

legibility of the controlled increases control for the controlling (Scott)

-> one-directional

information is not power but a precursor for enacting control

if your work is increasing algorithmic accuracy, who's control do you amplify?

seeing like a state/platform; influenceindustry.org; github; tracing apps; facial rec; …

Thread model 1

expect deployment before completion

simplicity fails to capture the task

goodhart’s law

human-specific classification boundary

short term

worst case: controller ignorance

Thread model 2

it actually works really well

controller group generates distribution

but group = subset of population / has weird preferences

-> lock in, stagnation, fragility ?

tax fraud detection; recidivism, Amazon hiring, skin cancer detection, Google facial rec;...

Thread model 1

prevent worst case

mostly don’t accelerate

use applications/production pipeline
to test best longterm procedures

Thread model 2

what does the world look like where
this didn’t happen / went ok?

agi is not an obvious end point

selection of control group /
accountability mechanisms

process > outcome

---

what procedure gives us *the best guarantee of convergence*, on robust or safe algorithm?

longtermist? because you attack risk production + predictions are less relevant

complaints about AI are often about inadequacy rather than about how it got decided to build algorithm x in the first place

safer design-deployment piple-line

---

status quo can work fine: google maps

status quo can work badly: tax fraud detection

at best: voluntary self-assessment reports (e.g Weidigner et al. 2021)

automation as standardisation: a resilience - fairness trade-off?

pol.is

Consequences of this framework

- △ between longtermist and shortermist AI risk reduction is small
  - more attention on: lock-in/stagnation
  - less attention on: extinction via AI/Super-AI/AGI
  - target optimal convergence process > optimal outcome

Consequences

‘solving’ alignment should not be the goal: coding ethics => lock-in

risks from malfunctioning AI < risks from functional AI

obedient AI does not solve the multi-agent alignment problem

politisise AI!

sig change going to correlate with change in control

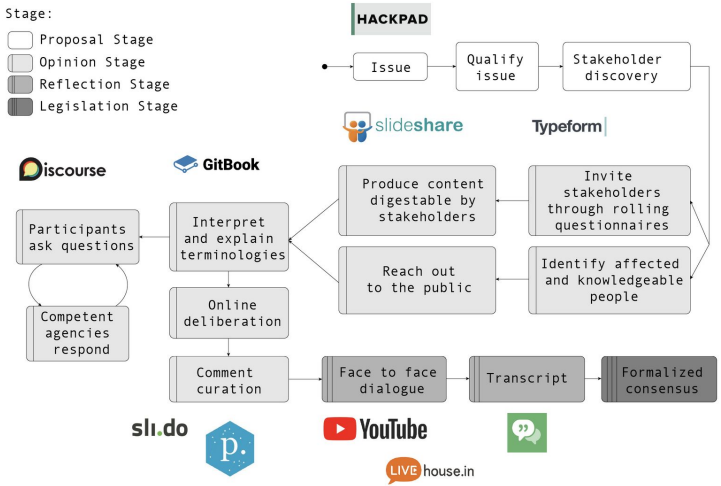
(xr to build explicit theory of power)

Reading

- Case Study ‘The Simple but Ingenious System Taiwan Uses to Crowdsource Its Laws’. MIT Technology Review.https://www.technologyreview.com/2018/08/21/240284/the-simple-but-ingenious-system-taiwan-uses-to-crowdsource-its-laws/ (May 2, 2022).
- Theory Landemore 2017, ISBN 9780691176390
- Tool Koster et al. 2022 https://doi.org/10.48550/arXiv.2201.11441 (investment & redistribution)

Scholars/Activists

- Kate Crawford,
- Alexa Koenig, - open source intelligence to be used in courts
- Audrey Tang, - vTaiwan
- Matthias Spielkamp, - algorithm watch
- Lilith Wittmann,
- Ruha Benjamin,
- Luke Kemp, Charlotte Siegman...



Orgs

read|design|databases|experiment

- explicit internal procedures
- Collate process features that generated failure
- design processes
  - run expt
  - iterate
  - document document document

R: e.g. Participedia.net / OECD. 2020

C -> decision w high uncertainty & value judgement & expertise, e.g. grant giving

P => limits, questions, outcome measurements, database on standadisation

Individuals

read|support|participate

- bellingcat
- chaos computer club
- data-altruism
- algorithm watch
- open knowledge foundation
- pol.is
- fragdenStaat + abgeordnetenwatch

Orgs

read|design|databases|experiment

- explicit internal procedures
- collate process features that generated failure
- design processes
  - run expt
  - iterate
  - document document document
- design tools
- fix identified problems
- G: controller becomes more transparent to the controlled
- fund

R: e.g. Participedia.net / OECD. 2020

C -> decision w high uncertainty & value judgement & expertise, e.g. grant giving

P => limits, questions, outcome measurements, database on standadisation

E.g. Rask, Mikko. 2013. 'The Tragedy of Citizen Deliberation. *https://doi.org/10.1080/09537325.2012.751012*

https://binary-butterfly.de/artikel/opendata-bisschen-prototyp-und-das-wars-dann/

Contact

carla.cremer@queens.ox.ac.uk

https://carlacremer.github.io/

@magdalenaadomeid