Moral Artificial Intelligence

Jean-François Bonnefon

Artificial and Natural Intelligence Toulouse Institute

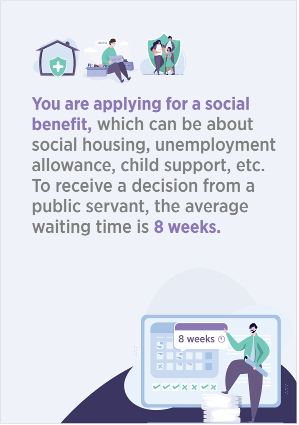
Morality in the age of intelligent machines

- Machine culture
  Nature Human Behaviour (2024)
- The moral psychology of AI
  Annual Review of Psychology (2024)
- Research on AI is reshaping our definition of morality
  Psychological Inquiry (2023)
- Moral AI and machine puritanism
  Brain & Behavioural Sciences (2023)
- Bad machines corrupt good morals
  Nature Human Behaviour (2021)
- Machine thinking, fast and slow
  Trends in Cognitive Science (2020)

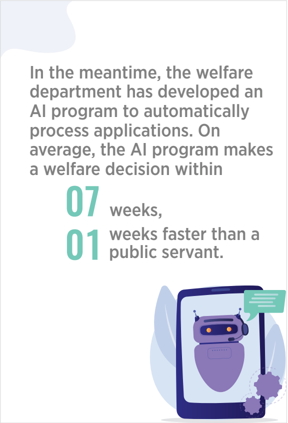
Behavioral work

- Alignment
  How do people want machines to make moral decisions?
- Scoring
  How do people react about being judged by machines?
- Cooperation
  Why and to what extent are people prosocial to machines?
- Disruptions
  How do machines transform our moral interactions?

Example of alignment: Welfare AI



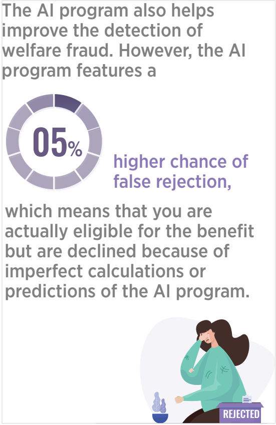
You are applying for a social benefit, which can be about social housing, unemployment allowance, child support, etc. To receive a decision from a public servant, the average waiting time is **8 weeks.**



In the meantime, the welfare department has developed an AI program to automatically process applications. On average, the AI program makes a welfare decision within

07 weeks,

01 weeks faster than a public servant.



The AI program also helps improve the detection of welfare fraud. However, the AI program features a

05% higher chance of false rejection,

which means that you are actually eligible for the benefit but are declined because of imperfect calculations or predictions of the AI program.

Example of alignment: Welfare AI

- At population level, preference curves are deceptively clear
  People trade 3-5 pp rejections for a 1 week speed gain
- But latent profile analysis reveals strong heterogeneity
  Vulnerable individuals hate rejections and don't care for speed
- Non-claimants don't understand preferences of claimants
  Claimants understand preferences of non-claimants
- Beware of paternalistic alignment
  We need vulnerable stakeholders engagement

Example of scoring: Moral uniqueness

- AI can be used to give moral scores to people
  From sexism on social media to full-blown social credit scores
- Acceptability of these scores is a function of their accuracy
  People don't want to be mischaracterized
- People believe scores will be inaccurate for uncommon moral profiles
  (Who knows if that's true)
- And they overestimate how uncommon their own profile is
  Hence low acceptability of moral scores, based on doubtful premises

Example of cooperation: The machine penalty

- People show non-zero prosociality to machines in incentivized games
  But not human-level prosociality: the ‘machine penalty’
- Anthropomorphizing machines is not a good fix
  Ineffective and/or ethically problematic
- But social norms are in flux and may erase the machine penalty
  Prosociality to machines is starting to signal prosociality to humans
- And the process can be accelerated by traditional tools
  Peer rewards, peer punishment, and their combination

Example of disruption: lie detection

Example of disruption: lie detection

- We are in a low accusation social equilibrium
  People lie but do not accuse others of lying
- Partly because of the (social) liability for false accusations
  Plus the fact we are bad at detecting lies
- But what if people could use LLMs to detect lies?
  How many would, and with each consequences on accusations?
- We investigated in an incentivized lie detection game
  Only 30% early adopters, but 4x effect on accusations

Moral Artificial Intelligence

Jean-François Bonnefon

Artificial and Natural Intelligence Toulouse Institute