A Bayesian Multilevel Analysis of Belief Alignment Effect Predicting Human Moral Intuitions of Artificial Intelligence Judgements

Yuxin Liu[1,2] and Adam Moore[1]
[1]Department of Psychology, School of Philosophy, Psychology and Language Sciences
[2]Centre for Technomoral Futures, Edinburgh Futures Institute

Introduction

Political Identity
Partisanship, Polarisation, Motivated Info Processing

Moral Judgements
Intuitions, Heuristics, Biases

HCI and Perceptions of AI
Inconsistent Evidence

Do people view AI as a neutral external viewpoint?

Research Question

Do people hold strong moral intuitions about AI generally, or do their judgements about AI vary systematically with their underlying politico-moral intuitions regarding the domain where the AI is deployed?

Hypotheses

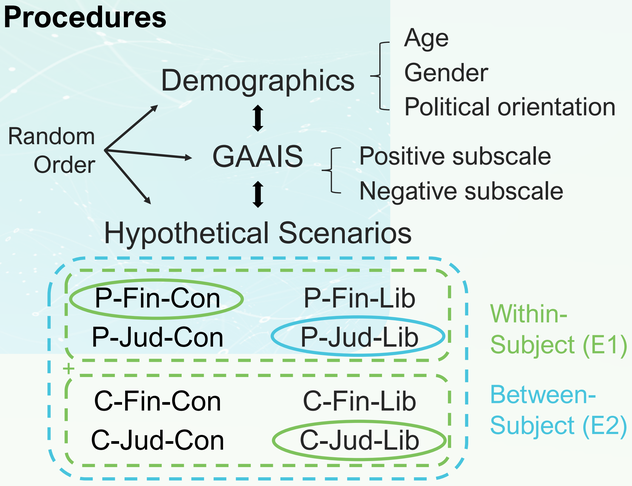
1. Belief alignment effects: when AI verdict aligns with politico-moral intuitions, participants will be more willing to act on its verdicts, trust it more, and perceive it as fairer;
2. Belief alignment effects will be stronger than/survive controlling for general AI attitudes;
3. Conservative/right-wing participants will show a stronger belief alignment effect than liberal/left-wing participants.

Methods

Participants

Native English-speaking adults in the UK
Recruited on Prolific Academic
E1: 202 (M_{age}= 36.7 yrs, SD_{age}= 13.36 yrs)
E2: 302 (M_{age}= 37.66 yrs, SD_{age}= 14.09 yrs)

Procedures



Key variables

- Main Predictors: Participant political orientation; Moral intuitive context
- Covariates: General attitudes towards AI; Age (nuisance)
- Outcome Variables: Willingness to act on AI verdict; Trust in AI; Perceived fairness of AI

Results & Discussion

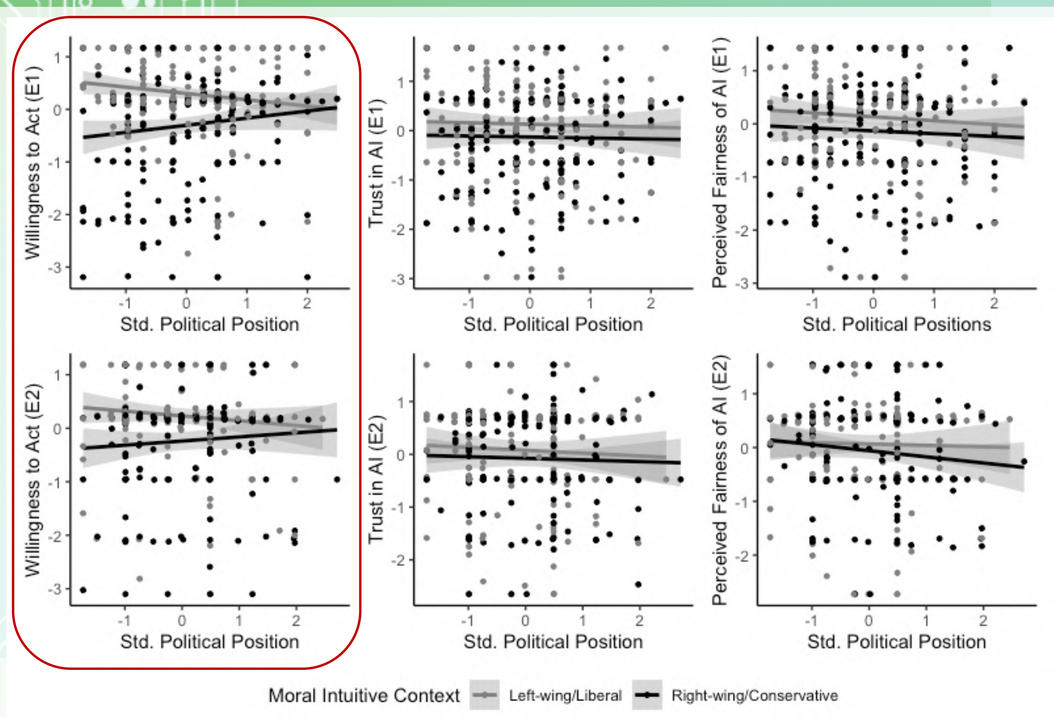


Figure 1: Belief alignment effect in E1 & E2 for Willingness to Act based on AI verdicts, but not for Trust in or Perceived Fairness of AI. Higher standardised scores on political position correspond to increasing conservatism.

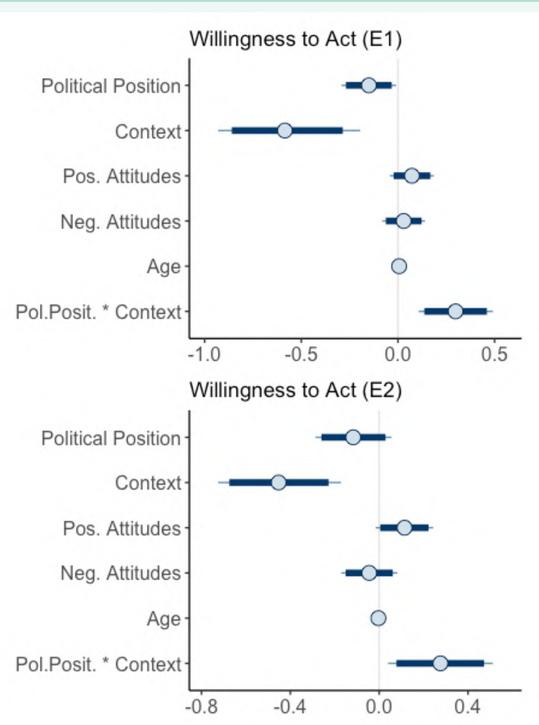


Figure 2: Posterior distributions with medians and 95% CIs for Willingness to Act based on AI verdicts in E1 & E2.

Judgements towards AI's detection of potential moral transgression seem to be at least partially driven by a compatibility between underlying values of AI verdicts and participants' own politico-moral beliefs, suggesting the malleability in framing the context of AI usage. This belief compatibility, however, did not increase trust or fairness perception of AI.

But… Measurement/sampling? Ideological coherence? Scenario complexity? Relevance of AI?