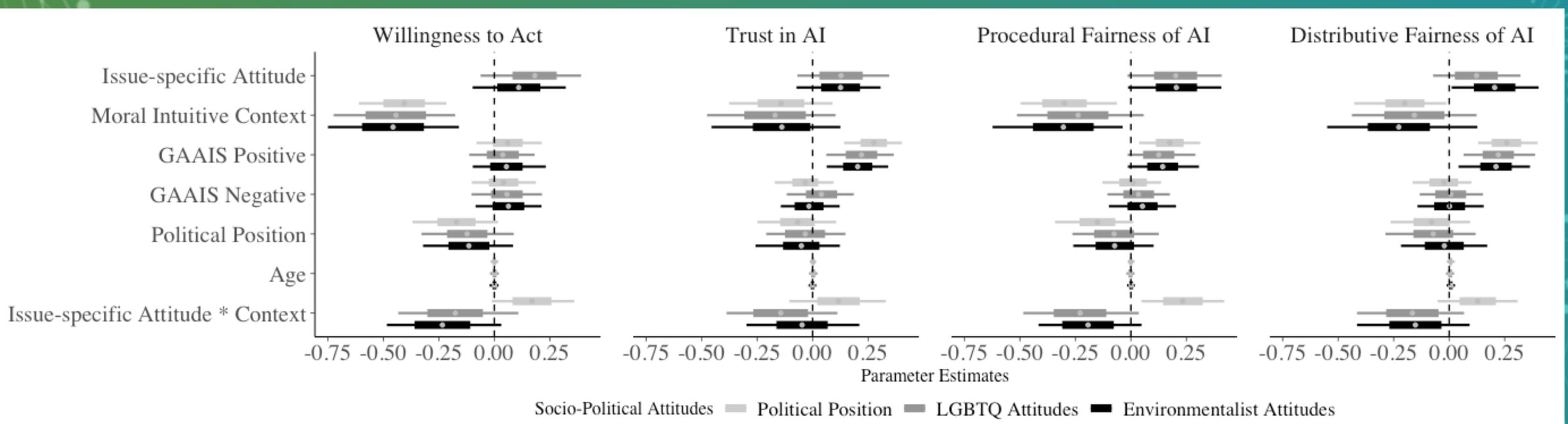
# Intuitive judgements towards AI verdicts of moral transgressions

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

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?




People's moral judgements towards AI verdicts of moral transgressions are constructed as functions of general positive AI attitudes, moral intuitive contexts of AI deployment, pre-existing politico-moral beliefs, and an alignment between the latter two.

## Introduction

Strong connection between **moral intuitions** and **political ideologies**

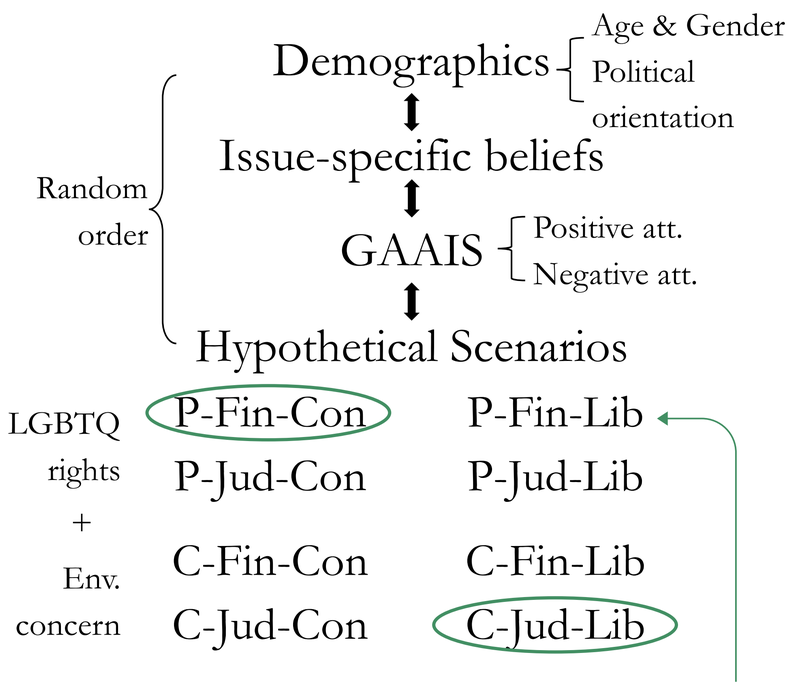
- Moral judgements influenced by intuition and heuristics
- Deeply-rooted sacred values protected against trade-off
- Political in-out group partisanship and motivated reasoning

→ Do people's beliefs about a given topic drive their acceptance/rejection of AI advice, or do people view AI suggestions as a kind of neutral external viewpoint that could cut through contentious issues?

→ **Hypothesis:** moral judgements towards AI verdicts are driven by an alignment between people's existing ideologies and underlying moral intuitive contexts of AI deployment over and above general AI attitudes

## Methods

**Participants** 302 native English-speaking adults ($M_{age}$ = 36.83 yrs, $SD_{age}$ = 10.79 yrs) in the UK recruited on Prolific Academic




Example: "A banking oversight committee has been using an efficient and reliable artificial intelligence system called Analytic Intellect to analyse loan application outcome patterns. The AI detected that a particular loan manager has been anomalously more likely to reject mortgage loan requests submitted by same-sex couples."

Outcome Variables:
- Willingness to act on AI verdict
- Trust in AI
- Procedural Fairness
- Distributive fairness

## Result Highlights

Willingness to act, trust, and fairness perception ↑ with

- positive attitudes towards practical utilities of AI
- endorsement for LGBTQ rights and environmentalism

Willingness to act, trust, and fairness perception ↓ with

- conservative AI contexts
- misalignment between issue-specific attitude and contexts of AI deployment

No influence from negative dystopian concerns of AI

Little influence from participant overall political orientations

Judgements towards AI are under the impact of both positive AI attitudes and a belief (mis)alignment effect, suggesting a level of malleability and context dependency influenced by pre-existing attitudes towards AI and an alignment between contexts and politico-moral beliefs.