Moral Intuitions Regarding the Use of Artificial Intelligence

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

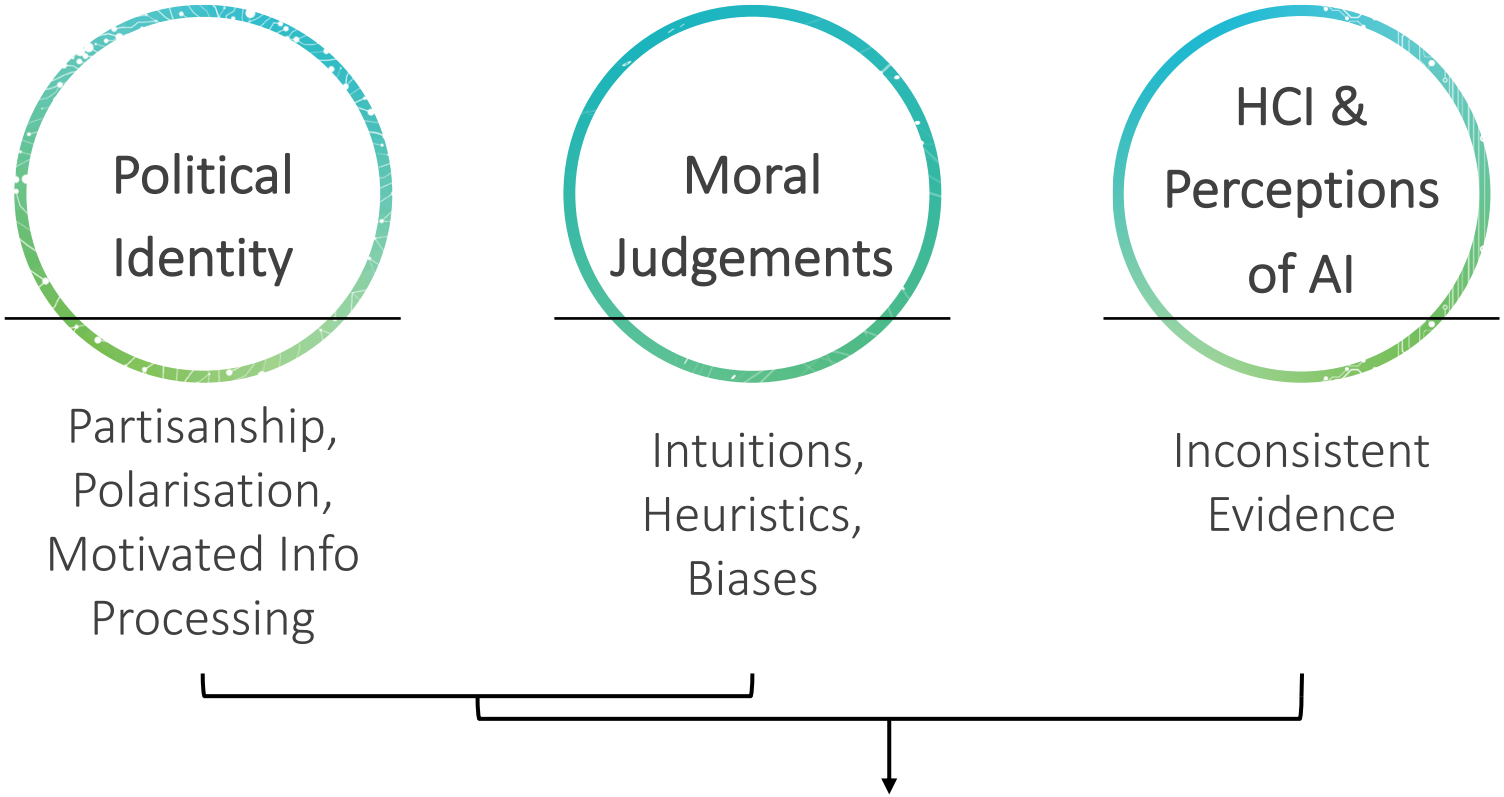
The 19th BPS Cognitive Section conference (*CogSec '22*)
September 6-9, 2022; Brighton, United Kingdom

Centre for **Technomoral Futures**

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"

Background



Will people view AI as neutral external third party that could potentially cut through divisive issues, or will their intuitions/beliefs about a given topic drive their judgements of AI advice?

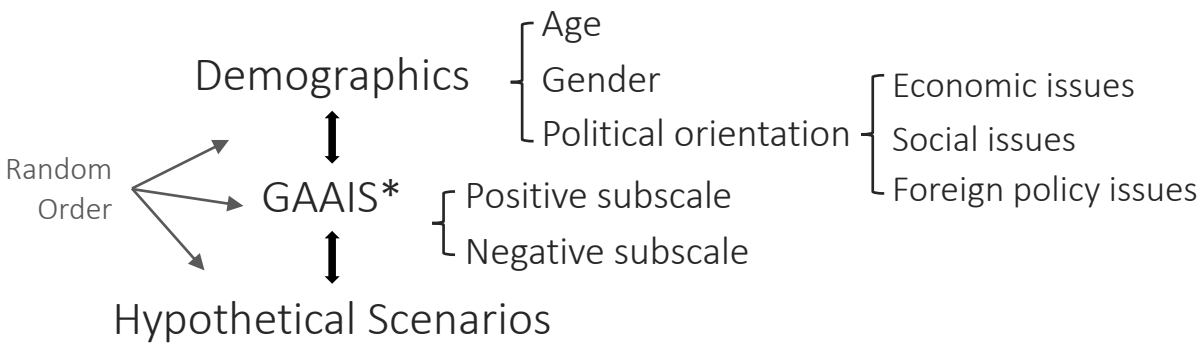
Research Question

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



*Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports, 1*, 100014.

Methods

Demographics

GAAIS

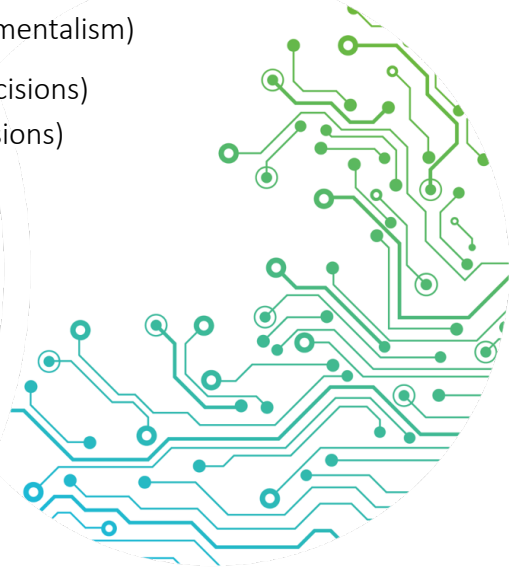
Hypothetical Scenarios

P-Fin-Con	P-Fin-Lib
P-Jud-Con	P-Jud-Lib
C-Fin-Con	C-Fin-Lib
C-Jud-Con	C-Jud-Lib

Person-centred (LGBTQ+ rights)
Cause-centred (Environmentalism)

Financial (banking decisions)
Judicial (judicial decisions)

Left-wing/Liberal
Right-wing/Conservative



Example

Financial

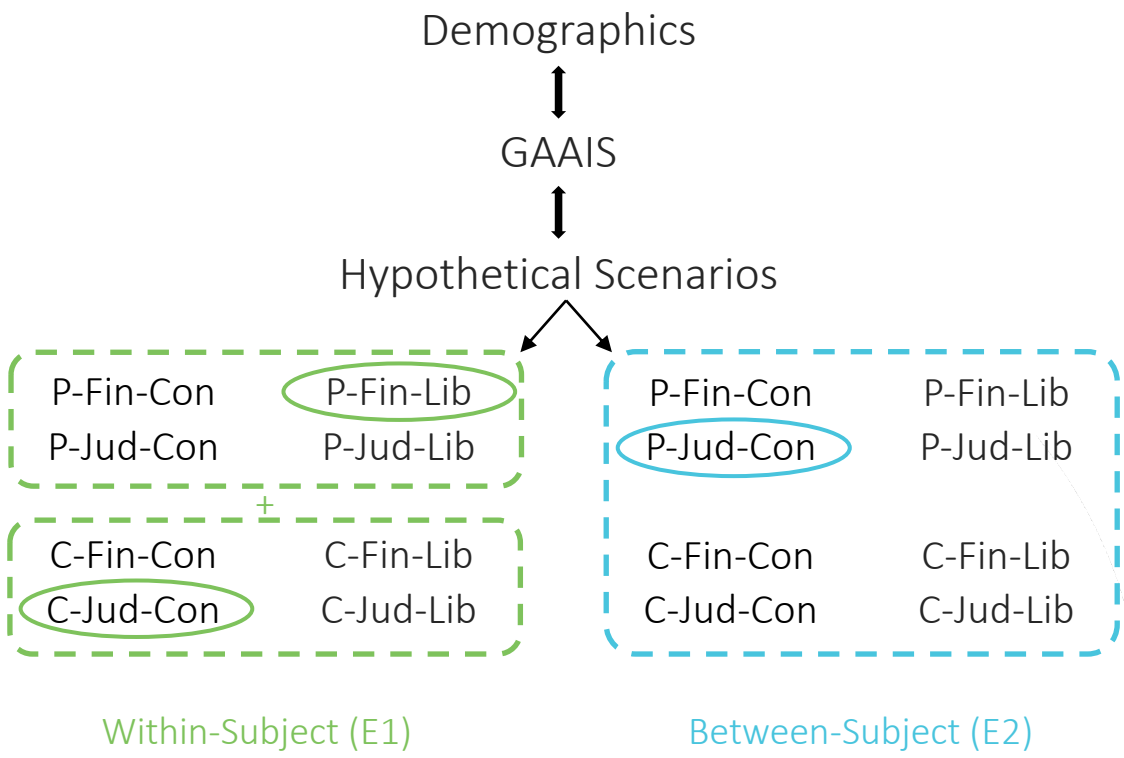
"A *banking oversight committee* has been using an efficient and reliable artificial intelligence system called Analytic Intellect to analyse loan application outcome patterns. The AI detected that a particular loan manager has been anomalously more likely to reject mortgage loan requests submitted by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

Person-centred

Launch an investigation into this loan manager

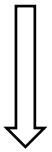
Left-wing/Liberal context

Methods


Demographics
GAAIS
Hypothetical Scenarios
P-Fin-Con
P-Fin-Lib
P-Jud-Con
P-Jud-Lib
+
C-Fin-Con
C-Fin-Lib
C-Jud-Con
C-Jud-Lib
Within-Subject (E1)
P-Fin-Con
P-Fin-Lib
P-Jud-Con
P-Jud-Lib
C-Fin-Con
C-Fin-Lib
C-Jud-Con
C-Jud-Lib
Between-Subject (E2)

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"

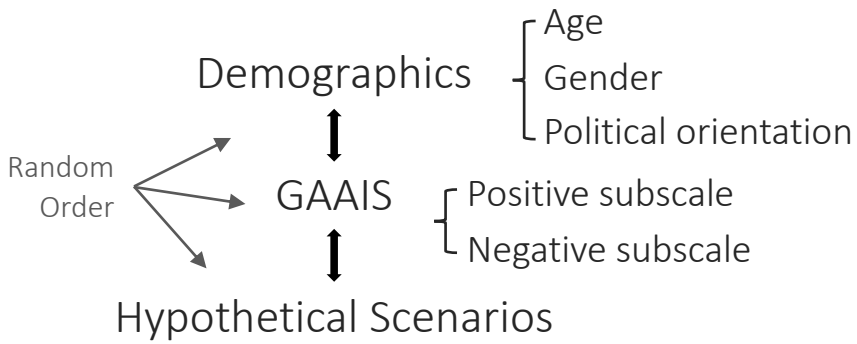


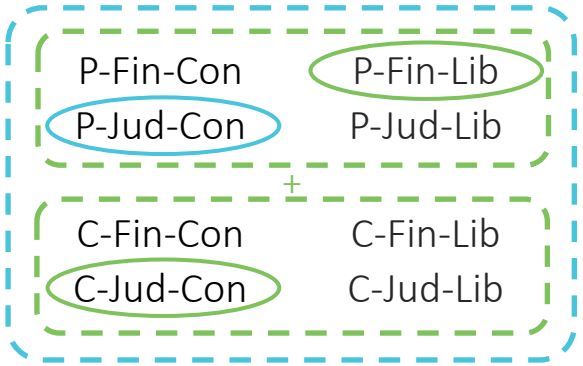
Scenario Responses

- Based on the AI's recommendation, I think that this person in the scenario should be investigated.
- I trust the AI's judgement in this case.
- I believe that the AI is being fair in this case.

1 = *strongly disagree*, 5 = *strongly agree*

Methods



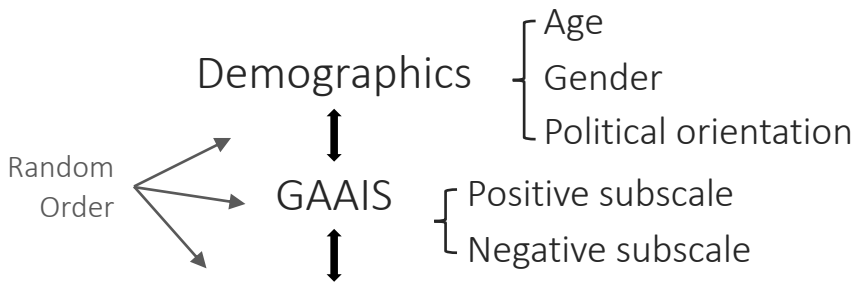


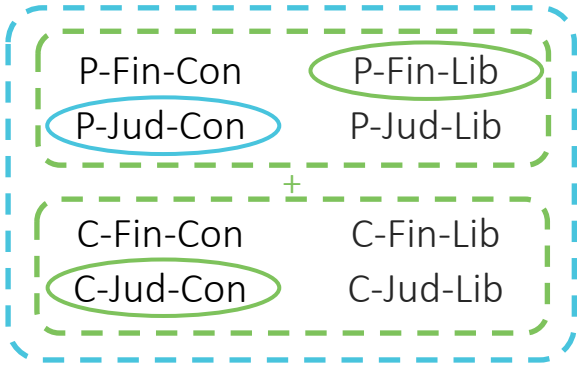
Within-Subject (E1)

Between-Subject (E2)

- Willingness to act on AI verdict
- Trust in AI
- Perceived fairness of AI

Methods





Within-Subject (E1)

Between-Subject (E2)

Bayesian multilevel regression

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

Methods

Participants: Native English-speaking adults in the UK recruited on Prolific Academic

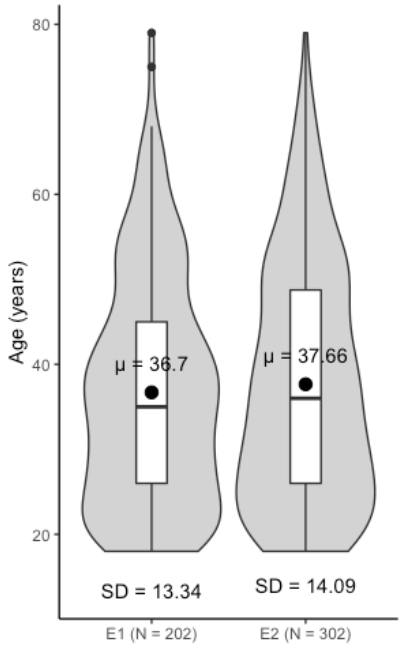


Figure 1: Violin plots for distributions of participant age in E1 & E2.

Methods

Participants: Native English-speaking adults in the UK recruited on Prolific Academic

Table 1: Descriptive summaries of measured variables in E1 & E2.

	Experiment 1 (within-subjects)			Experiment 2 (between-subjects)		
	Mean (SD)	Median	Range	Mean (SD)	Median	Range
Political Positions (1 = *Very Left/Liberal*, 7 = *Very Right/Conservative*)						
Economic Issues	3.39 (1.33)	3.00	6.00	3.47 (1.34)	4.00	6.00
Social Issues	3.15 (1.38)	3.00	6.00	3.16 (1.32)	3.00	6.00
Foreign Policy Issues	3.37 (1.34)	4.00	6.00	3.39 (1.40)	4.00	6.00
Mean Political Position	3.30 (1.25)	3.33	5.67	3.34 (1.25)	3.33	6.00
General Attitudes Towards AI (1 = Negative Attitudes, 5 = Positive Attitudes)						
Positive Subscale	3.33 (0.60)	3.33	2.75	3.30 (0.60)	3.33	3.50
Negative Subscale	2.97 (0.65)	3.00	3.25	3.04 (0.69)	3.12	3.75
Responses to Scenarios (1 = *Strongly Disagree*, 5 = *Strongly Agree*)						
Willingness To Act	3.93 (0.92)	4.07	4.00	3.89 (0.93)	4.06	4.00
Trust	3.56 (0.86)	3.62	4.00	3.44 (0.92)	3.69	4.00
Perceived Fairness	3.67 (0.93)	3.94	4.00	3.56 (0.94)	3.78	4.00

Note. For meaningful interpretations, descriptive statistics are presented in original scales of measurement.

Results

Table 2: Bayesian Pearson's zero-order correlations and their 95% HDIs between main variables in E1 (the lower diagonal) & E2 (the upper diagonal).

E1 \ E2	Political Positions	Positive Attitudes	Negative Attitudes	Willingness to Act	Trust	Perceived Fairness
Political Positions	1	-0.13** [-0.22, -0.04]	-0.13* [-0.23, -0.05]	-0.02 [-0.11, 0.07]	-0.04 [-0.14, 0.04]	-0.07 [-0.16, 0.02]
Positive Attitudes	-0.06 [-0.15, 0.01]	1	0.50*** [0.44, 0.58]	0.11* [0.02, 0.20]	0.05 [-0.04, 0.14]	0.07 [-0.02, 0.16]
Negative Attitudes	0.05 [-0.03, 0.13]	0.51*** [0.45, 0.56]	1	0.03 [-0.07, 0.11]	-0.01 [-0.10, 0.08]	-0.00 [-0.10, 0.08]
Willingness to Act	0.01 [-0.07, 0.08]	0.07 [0.00, 0.16]	0.06 [-0.03, 0.13]	1	0.35*** [0.27, 0.43]	0.36*** [0.28, 0.43]
Trust	-0.02 [-0.10, 0.06]	0.20*** [0.13, 0.28]	0.14** [0.07, 0.22]	0.31*** [0.24, 0.38]	1	0.63*** [0.57, 0.68]
Perceived Fairness	-0.06 [-0.14, 0.02]	0.21*** [0.13, 0.28]	0.10* [0.02, 0.18]	0.36*** [0.29, 0.43]	0.62*** [0.56, 0.66]	1

Note. Probability of direction (pd) represents the portion of the posterior distribution in the same direction of effect as the median (Makowski et al., 2019); *** pd > 99.95%, ** pd > 99.5%, * pd > 97.5%. Negative attitudes are reverse-coded.

Results

Table 3: Full summaries of Bayesian regression fixed effects coefficients for E1 & E2.

	Willingness to Act		Trust		Fairness Perception	
Experiment 1	Mean [95% HDI]	SD	Mean [95% HDI]	SD	Mean [95% HDI]	SD
Intercept	0.07 [-1.01, 1.13]	0.50	0.17 [-0.74, 1.08]	0.42	0.16 [-0.80, 1.09]	0.43
Political Position	**-0.15 [-0.29, -0.01]**	0.07	-0.04 [-0.19, 0.11]	0.08	-0.09 [-0.24, 0.06]	0.07
Context	**-0.58 [-0.93, -0.20]**	0.18	-0.25 [-0.52, 0.03]	0.14	-0.26 [-0.53, 0.02]	0.14
Positive Attitudes	0.07 [-0.04, 0.19]	0.06	**0.16 [0.04, 0.28]**	0.06	**0.21 [0.09, 0.33]**	0.06
Negative Attitudes	0.03 [-0.08, 0.14]	0.06	0.07 [-0.05, 0.19]	0.06	0.01 [-0.11, 0.13]	0.06
Age	0.01 [0.00, 0.01]	0.00	0.00 [-0.01, 0.01]	0.00	0.01[-0.01, 0.01]	0.00
Political Position * Context Interaction	**0.30 [0.11, 0.49]**	0.10	0.06 [-0.13, 0.26]	0.10	0.08 [-0.10, 0.27]	0.09
	Willingness to Act		Trust		Fairness Perception	
Experiment 2	Mean [95% HDI]	SD	Mean [95% HDI]	SD	Mean [95% HDI]	SD
Intercept	0.35 [-0.80, 1.48]	0.54	0.04 [-0.92, 1.00]	0.44	-0.05 [-0.99, 0.91]	0.44
Political Position	-0.12 [-0.29, 0.06]	0.09	-0.11 [-0.29, 0.07]	0.09	-0.08 [-0.26, 0.10]	0.09
Context	**-0.45 [-0.72, -0.17]**	0.14	-0.14 [-0.43, 0.16]	0.15	-0.10 [-0.46, 0.25]	0.18
Positive Attitudes	0.11 [-0.02, 0.24]	0.07	0.09 [-0.05, 0.23]	0.07	0.13 [-0.01, 0.26]	0.07
Negative Attitudes	-0.04 [-0.17, 0.08]	0.06	-0.05 [-0.18, 0.08]	0.07	-0.06 [-0.19, 0.07]	0.07
Age	0.00 [-0.01, 0.00]	0.00	0.00 [-0.01, 0.01]	0.00	0.00 [-0.01, 0.01]	0.00
Political Position * Context Interaction	**0.28 [0.04, 0.51]**	0.12	0.14 [-0.11, 0.38]	0.13	0.02 [-0.23, 0.26]	0.12

Note. Model converged successfully with split R-hat = 1 for all estimated parameters. Context is a binary variable with liberal/left-wing direction as the reference level. Negative attitudes are reverse-coded. Bold emphasises $0 \notin 95\%$ *HDI*.

Results

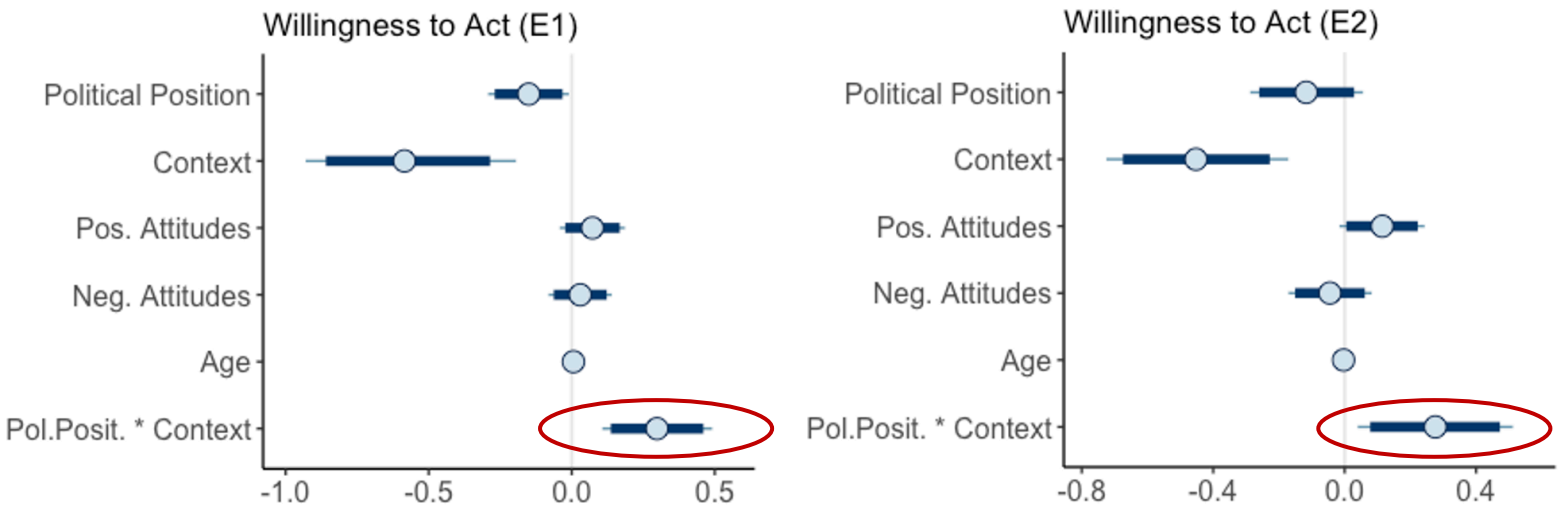


Figure 2: Posterior distributions with medians and 95% CIs for Willingness to Act based on AI verdicts in E1 & E2.

Results

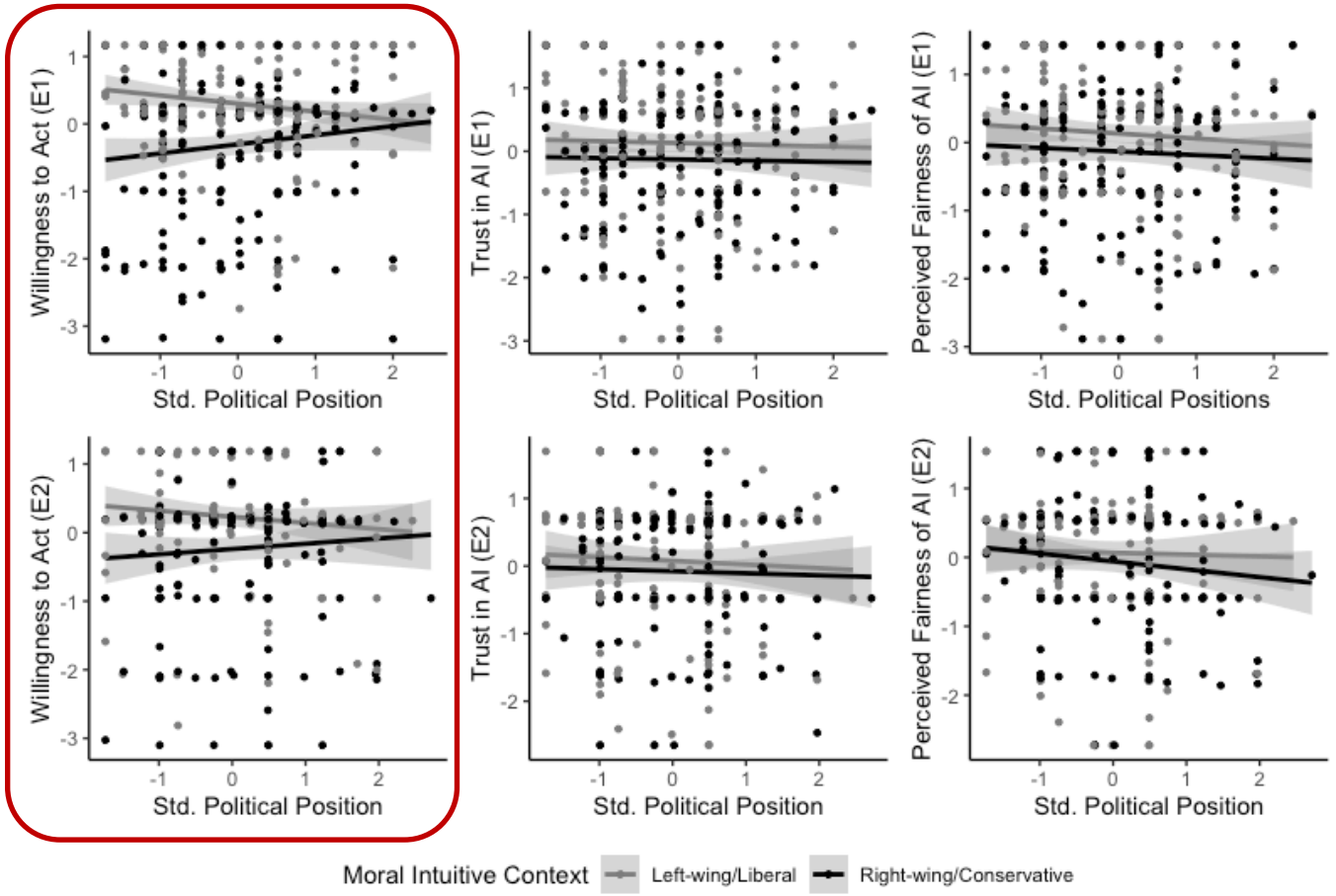


Figure 3: Belief alignment effect in E1 & E2 for Willingness to Act based on AI verdicts, but not for Trust in or Perceived Fairness of AI. Higher standardised scores on political position correspond to increasing conservatism.

Discussion

Generally less willing to act on verdicts of wrongdoing in conservative contexts vs. liberal ones

Willingness to act on AI advice was predominantly driven by a belief-alignment effect

- i.e., whether the AI's recommendation aligned with pre-existing politico-moral intuitions cued by the scenario context
- consistent with motivated social cognition needs
- trumped general AI attitudes - people likely have weak to no moral intuitions about AI itself

Belief-alignment did not increase trust/fairness perception of AI

- Disassociation between willingness to act on AI advice and judgements of trustworthiness/fairness of AI
- Distributive fairness vs. procedural justice?

…but

- Measurement/sampling?
- Scenario complexity?
- Ideological coherence?
- Relevance of AI?

Full paper link:

Liu, Y. and Moore, A. (2022). A Bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgements. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 2116–2125. https://escholarship.org/uc/item/3v79704h

Centre for Technomoral Futures

https://www.technomoralfutures.uk/phd-students/yuxin-liu

yliu3310@ed.ac.uk

@_yuxinl_