Received: 20 October 2024 | Accepted: 12 May 2025

DOI: 10.1111/bjso.12908

ARTICLE

Intuitive judgements towards artificial intelligence verdicts of moral transgressions

Yuxin Liu[1,2] | **Adam Moore**[1]

[1]School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, UK

[2]Centre for Technomoral Futures, Edinburgh Futures Institute, The University of Edinburgh, Edinburgh, UK

Correspondence
Yuxin Liu, Edinburgh Futures Institute, Centre for Technomoral Futures, The University of Edinburgh, R3.50, 1 Lauriston Pl, Edinburgh EH3 9EF, UK.
Email: yuxin.liu@ed.ac.uk

Funding information
University of Edinburgh; Baillie Gifford

Abstract

Automated decision-making systems have become increasingly prevalent in morally salient domains of services, introducing ethically significant consequences. In three pre-registered studies ($N = 804$), we experimentally investigated whether people's judgements of AI decisions are impacted by a belief alignment with the underlying politically salient context of AI deployment over and above any general attitudes towards AI people might hold. Participants read conservative- or liberal-framed vignettes of AI-detected statistical anomalies as a proxy for potential human prejudice in the contexts of LGBTQ+ rights and environmental protection, and responded to willingness to act on the AI verdicts, trust in AI, and perception of procedural fairness and distributive fairness of AI. Our results reveal that people's willingness to act, and judgements of trust and fairness seem to be constructed as a function of general attitudes of positivity towards AI, the moral intuitive context of AI deployment, pre-existing politico-moral beliefs, and a compatibility between the latter two. The implication is that judgements towards AI are shaped by both the belief alignment effect and general AI attitudes, suggesting a level of malleability and context dependency that challenges the potential role of AI serving as an effective mediator in morally complex situations.

KEYWORDS

automated decision-making, belief alignment, human–AI interaction, moral intuitions, motivated reasoning

INTRODUCTION

Contrary to the idiom, facts do not speak for themselves. Research in social and political psychology, particularly on identity politics and in-group/out-group partisanship (Bartels, 2002; Cohen, 2003), shows identical incoming information can be processed in distinctively different ways, that is individuals tend to accept/reject incoming information and react to the behaviour of others as a function of (mis)alignment with their existing ideology/worldview (Cook & Lewandowsky, 2016; Cram et al., 2018; Drummond & Fischhoff, 2017; Gaines et al., 2007; Hameleers & van der Meer, 2020; Jern et al., 2014; Moore et al., 2021). Notably, this kind of selective distortion can occur regardless of factual accuracy, or even at the expense of it, for a host of political/politicized issues such as climate change (McCright & Dunlap, 2011), immigration and asylum seekers (Glinitzer et al., 2021) and Brexit (Hobolt et al., 2021). Such selective information processing has been more often explained by accounts of motivated reasoning (Flynn et al., 2017; Jost, 2017; Jost et al., 2003, 2017; Jost & Amodio, 2012; Jost & Krochik, 2014; Kahan, 2016a, 2016b; Krochik & Jost, 2011; Moore et al., 2021; Taber & Lodge, 2006) than by accounts of effortful rejection of information (Pennycook & Rand, 2019; Roozenbeek & van der Linden, 2019).

Whilst these self-perpetuating belief systems and associated ideological/affective polarization (Geschke et al., 2019; Iyengar et al., 2019; van Baar & FeldmanHall, 2021) present inherent challenges, they are further complicated by technological development that raises concerns over fairness/biases, mis/disinformation, opacity/explainability and more (Weidinger et al., 2022). These low-trust sociopolitical environments with entrenched belief polarization, nonetheless, also present opportunities for third-party agents to provide guidance – could AI effectively fulfil this role? The term 'AI' here refers to a broad umbrella of decision-assisting technologies. For example, earlier studies in human–computer interaction tend to focus on statistical forecasting algorithms and the resulting human preference for, or prejudice against, algorithmic predictions (Dietvorst et al., 2015; Logg et al., 2019). More recently, advancements in large language models have shifted attention towards examining the perception of AI-generated human-like utterances and their persuasive impact on users' judgements and beliefs (Costello et al., 2024). In the adjacent field of machine ethics that seeks to embed moral capacity in robots, machines and AI, several theoretical proposals have emerged for building artificial moral advisors using various philosophical frameworks and design approaches (Giubilini & Savulescu, 2018; Lara & Deckers, 2020; but see Liu et al., 2022).

These developments raise a critical question: to what extent do individuals' pre-existing intuitions and beliefs about a given topic shape their judgements towards AI-generated output? This is a key concern because, if motivated reasoning drives user evaluation of AI beyond general attitudes towards AI, then the potential for AI systems to serve as mediators in morally/politically contentious situations will be significantly undermined. More concerning is the risk of asymmetric acceptance of AI-generated content, where individuals dismiss AI as biased/unreliable when it challenges their pre-existing beliefs but accept it as independent/trustworthy when it aligns with them. Given the growing prevalence of AI operating in morally salient contexts, it is crucial to understand how people assess AI as a mediator in complex moral decision-making processes.

Extensive research has been conducted on human–AI/computer interaction, and more recently, on the moral psychology of AI (Bonnefon et al., 2024; Ladak et al., 2023). Earlier studies comparing humans and statistical algorithms have often demonstrated an algorithm aversion effect – prejudicial discounting of algorithmic decisions or a preference for human decisions (Dietvorst et al., 2015; Jauernig et al., 2022; Longoni et al., 2019; Morewedge, 2022; Önkal et al., 2009; Prahl & van Swol, 2017; Zhang et al., 2021), though fewer studies show the opposite – an algorithm appreciation (Logg et al., 2019; see Burton et al., 2020; Mahmud et al., 2022 for systematic reviews). Beyond preferences, abundant research has explored moral norms, fairness, responsibility and accountability in judgement and decision-making involving AI (Araujo et al., 2020; Banks, 2020; Bonnefon et al., 2016; Hong et al., 2020; Kahn et al., 2012; Malle et al., 2015, 2019; Shank et al., 2019, 2021; Shank & DeSanti, 2018; Shariff et al., 2017), though definitive conclusions remain elusive.

These studies tend to focus on elements related to the internal design of algorithms, individual differences in psychological features, task characteristics and higher-level (e.g. cultural/societal) factors (Mahmud et al., 2022). Task characteristics, for example, are crucial in AI perception: individuals display greater trust in, and perceived fairness of, AI, for tasks perceived as objective/quantifiable, as opposed to subjective/associated with human affective abilities (Castelo et al., 2019; Lee, 2018). Though many studies considered the influence of familiarity or experience with AI, the role of general attitudes towards AI is often overlooked, and more importantly, so is whether the AI output agrees with participants' pre-existing socio-moral beliefs. The potential effects of motivated reasoning in human–AI interactions are particularly relevant because, unlike static features of the technology/individual user, pre-existing beliefs and ideologies can be amplified through repeated interactions with AI systems and diffused across social networks, reinforcing and spreading biased judgements over time (Glickman & Sharot, 2025).

The current research

Given the tendency to evaluate incoming information based on its compatibility with existing beliefs, does this pattern extend to information coming from an AI? And if so, does this pattern hold beyond any intuitions that people might have about AI itself? Using vignettes of narrow AI deployment, we aim to investigate whether judgements towards AI are impacted by belief alignment with the AI's output in its deployment context over and above general AI attitudes. Though we do control for general/broader AI attitudes, focusing on a narrow AI with a specific, pattern-detection function allows us to target this (mis)alignment somewhat separately from potential assumptions about more complex, human-like AI.

To reliably elicit polarized beliefs independent of general AI attitudes, these scenarios involve politically salient contexts (LGBTQ+ rights and environmental protection), which allow us to examine whether AI detection of potential bias in human behaviour is considered sufficient evidence for transgressions that warrant investigative actions. We experimentally manipulate the (mis)alignment between AI recommendation and participants' pre-existing politico-moral intuitions for the underlying context, aiming to prompt a compatibility or conflict between them. The rationale is that if judgements towards AI output are contingent on pre-existing intuitions related to the context, then belief alignment should lead to a greater acceptance of AI verdicts, and vice versa, notwithstanding any general intuitions people might have about AI.

Additional to willingness to act, judgements of trust and perception of fairness are also two prevalent dimensions of moral judgements and are both linked to reactions to decision outcomes (Bianchi et al., 2015; Skitka & Mullen, 2002; Tyler & Degoey, 1996; Tyler & Smith, 1999). They are not only frequently studied in judgements towards different types of authorities/experts (de Cremer & Tyler, 2007; Promberger & Baron, 2006), but also in human–machine interactions (Castelo et al., 2019; Grgić-Hlača et al., 2018, 2022; Lee, 2018; Lee & Baykal, 2017; Lee & Rich, 2021; Malle & Ullman, 2021; Starke et al., 2022; Ullman & Malle, 2018, 2019). The majority of literature on fairness perceptions in algorithmic decisions (Starke et al., 2022) focuses on identifying algorithmic or human predictors of fairness perception of AI and testing the consequences of perceived fairness. Relatively more research has investigated trust(worthiness) of AI in a similar fashion (see Bach et al., 2022; Ueno et al., 2022 for systematic reviews). This includes examining technical design features and human factors that enhance or undermine trust in human–machine interaction, devising measurement of trust in machines (Malle & Ullman, 2021; Ullman & Malle, 2018, 2019), comparing trust in human/AI (Liang & Newell, 2022) and exploring relationships between trust and other variables (Choung et al., 2023). As such, we will also investigate whether trust and fairness perception are subject to context-based motivated reasoning in cases of AI/automated decisions.

Lastly, considerable research in political psychology reveals an ideological asymmetry, such that conservatives, compared to liberals, tend to score higher on variables such as cognitive rigidity and dogmatism, and lower on integrative complexity, cognitive reflection, and non-political measures of cognitive flexibility (Amodio et al., 2007; Jost, 2017; Zmigrod et al., 2020). Thus, albeit not the main objective, we expect to observe a greater belief (mis)alignment effect for conservatives, given that on average, conservatives are more susceptible to various cognitive and epistemic biases (Baron & Jost, 2019) and political conservatism may be, in part, a manifestation of motivated social cognition (Jost et al., 2003).

We conducted three pre-registered experiments (https://osf.io/7qjt3). In E1 (https://osf.io/wufpg) and E2 (https://osf.io/q82ec), we examined how willingness to act on AI verdicts of moral transgression, trust in AI and perceived fairness of AI varied with alignment between participants' overall political orientation and AI recommendations beyond general AI attitudes in contexts of LGBTQ+ rights and environmental concerns. In E3 (https://osf.io/wm29n), we used issue-specific measures to better capture the alignment in those contexts. Results consistently showed that belief alignment impacted judgements towards AI verdicts over and above general AI attitudes, suggesting a level of malleability and context dependency that challenges AI's potential mediating role in morally charged situations.

EXPERIMENT 1 AND 2

Given the well-documented tendency for motivated reasoning, judgements about AI outcomes may be predominantly driven by context-based belief alignment depending on the compatibility between AI verdicts of moral transgressions and pre-existing politico-moral beliefs. We predict: (1) increased willingness to act on AI recommendations, increased trust in AI and increased perception of fairness of AI when those AI recommendations align with participants' pre-existing moral intuitions, compared to when they do not align (i.e. a belief alignment effect), (2) belief alignment will occur over and above general AI attitudes and (3) conservatives showing stronger belief alignment than liberals.

We conducted two experiments using the same study design and materials: E1 was a repeated-measures design (multiple scenarios per participant) and E2 was between-subjects (one scenario per participant). A consistent pattern of results across the two experiments should increase our confidence in the effect of belief alignment.

METHODS

Participants

After excluding one participant from each experiment for failing the attention check, we had 201 (67 males and 131 females; M_{age} = 36.70 years, SD_{age} = 13.38 years) and 301 native English-speaking adult participants (109 males and 190 females; M_{age} = 37.68 years, SD_{age} = 14.11 years) in E1 and E2, respectively. Testing was conducted online via Qualtrics integrated into the crowdsourcing platform Prolific to recruit diverse, representative, attentive and naïve subjects (Palan & Schitter, 2018; Peer et al., 2017, 2021). Participants were compensated £0.84 for E1 and £0.59 for E2, and repeat participation was prevented via Prolific internal filtering.

Design and materials

We collected data on (1) basic demographics and political orientation, (2) general attitudes towards AI and (3) intuitive responses to hypothetical scenarios of AI verdicts described as AI-detected statistical anomalies indicative of moral transgressions. These three sections were presented in a random order.

Demographic information

We collected participants' age, gender and aspects of political orientation. To account for different underlying political attitudes associated with facets of conservatism (Crowson, 2009; Harnish et al., 2018; Pratto et al., 1994), we measured political positions via one question each on economic, social and foreign policy views ('Using the following scale, how left-wing/liberal or right-wing/conservative are you on economic/social/foreign policy issues?'; 1 = *very left-wing/liberal* to 7 = *very right-wing/conservative*).

General attitudes towards AI

We used the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2020, 2022) to measure overall sentiments towards AI in the general public. The GAAIS consists of a positive subscale – 12 items capturing the practical functionality and potential societal/personal benefits of AI applications (e.g. 'I am interested in using artificially intelligent systems in my daily life'; 1 = *strongly disagree* and 5 = *strongly agree*), and a negative subscale – eight items capturing dystopian concerns towards the presumed danger of AI (e.g. 'I think artificial intelligence is dangerous'; 1 = *strongly agree* and 5 = *strongly disagree*). Negative items were reverse-coded so that higher ratings on both subscales would indicate more positive general attitudes towards AI. We calculated subscale means separately as instructed, due to the lack of unidimensionality.

Hypothetical scenarios

We created eight scenarios where organizations use a reliable expert AI system to evaluate their everyday operations, and the AI system detects statistically anomalous decisions made by a human agent in the organization (e.g. Table 1). We described a statistical pattern-detection algorithm here, rather than a more topical generative AI, for simplicity and to minimize potential confounding variables such as perceptions of agency, autonomy and humanness. Statistical anomalies as an insinuation of potential bias allow for a more focused study design without implicit assumptions of more sophisticated AI systems. We return to this point in the General Discussion.

These scenarios represent a fully factorial 2 (Context: Left-wing/Liberal or Right-wing/Conservative moral intuitive direction) × 2 (Approve or Reject action taken by the human agent) × 2 (Financial or Judicial domain of the scenario) design. Context indicates an AI verdict compatible with either liberal or conservative moral intuitions (e.g. an AI flagging a judge for prejudice against same-sex couples aligns with left-wing/liberal intuitions that such discrimination is wrong and should be stopped). Approve/reject action indicates the human agent favouring or discriminating against a target. Domain indicates the superficial setting of the scenarios (financial bank or judicial court), which are nested in a person-centred (LGBTQ+ rights) and a cause-centred (environmental concerns) focus. All elements (context, action, domain and focus) are counterbalanced.

Participants responded to three statements following each scenario, with each statement presented on a continuous slider (1 = *strongly disagree* to 5 = *strongly agree*) with a midpoint default. Willingness to act on AI recommendations refers to participants' support for default interventions (e.g. an investigation) based solely on AI detection of possible prejudice ('Based on the AI's recommendation, I think that this

TABLE 1 Example hypothetical scenarios.

Left-wing/Liberal context	Right-wing/Conservative context
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*.	A leading technology company has partnered with the Ministry of Justice to develop and train an artificial intelligence named LEA (Legal Expert Assistant) to serve *judicial needs*. The main objective of this AI is to identify any statistical anomalies in civil judicial decisions, which would potentially be flagged for re-evaluation. When reviewing the results of environmental claims cases in the past year, LEA detected that a particular judge has been ruling *in favour of* claims against corporations in *pollution or environmental damage* cases at a significantly higher rate than average.

Note: Italics indicate domains, actions and foci for clarity; no text was italicized for the participants. Overall length of scenarios did not differ as a function of context alignment.

person in the scenario should be investigated'). Trust in AI refers to the extent to which participants perceive the AI judgement as trustworthy ('I trust the AI's judgement in this case'). Perceived fairness of AI refers to the extent to which they perceive the AI as fair and appropriate ('I believe that the AI is being fair in this case').

Procedures

After giving informed consent, participants completed the above-mentioned three sections in a random order. E1 participants received two pseudo-randomly selected scenarios from opposite factorial cells in each topical focus, and E2 participants were shown one randomly selected scenario out of the eight possibilities. After each scenario, participants responded to statements on willingness to act, trust and perceived fairness, one at a time on separate pages while the given scenario remained visible above each statement. For both experiments, all scenarios were approximately evenly presented across participants.

Statistical analysis plan

We opted for Bayesian statistical analysis (using *brms* package (v. 2.19.0); Bürkner, 2017, 2018) in RStudio (v. 4.3.0; R Core Team, 2023) to quantify support for our hypotheses of interest, rather than the (in) compatibility of the evidence with the null hypothesis (McElreath, 2015). Under the Bayesian framework, we computed zero-order correlations and multilevel multivariate multiple regression models, with the main pre-registered model containing the predictors of interest, covariates and a nuisance variable (see below).

All continuous variables were standardized prior to analysis. We averaged the three standardized political view items to obtain an overall measure of participant political position, where higher scores indicate increasing right-wing conservatism. Fixed effects of context, participant political orientation and the interaction of the two were entered into the models as main predictors of interest. We used a prior of β~normal (0, .25) for the intercept, and weakly regularizing priors of β~normal (0, 1) for context and political orientation, whose effects were not explicitly predicted. For the interaction term between those two, the prior was set at β~normal (.3, .15), indicating the hypothesized increases in willingness to act, trust and fairness perception as a function of belief alignment. Means of positive and (reverse-coded) negative subscales of GAAIS were entered as covariates of interest to account for participants' general views of AI unrelated to our scenario design. Both GAAIS subscales had a prior of β~normal (.2, .1) to indicate our expectation that general optimism towards AI would likely predict more positive judgements about AI verdicts, but to a lesser degree compared to belief alignment effect. Age was also included as a nuisance covariate to represent basic familiarity with AI, with a prior of β~normal (−.1, .1) to account for its small negative effect on perception of technologies (Schepman & Rodway, 2020, 2022). Unique idiosyncrasies within each item, topic and individual subject were modelled with random intercepts. All the above parameters and priors were used to simultaneously predict willingness to act on AI verdicts, trust in AI and perceived fairness of AI, thus controlling for correlations amongst them and generating unique predictive effects for each outcome.

Posterior distributions of regression parameters were derived by simulation using Markov chain Monte Carlo (MCMC) estimation (Betancourt, 2018; Bürkner, 2017, 2018; Gelman & Rubin, 1992). For all models, we sampled from four independent MCMC chains with 1000 burn-in samples and 15,000 sampling iterations per chain. All models converged (all $\widehat{R}$ *s* = 1.0; Brooks & Gelman, 1998; Gelman et al., 2013; Gelman & Rubin, 1992). Effect size uncertainty is computed as 95% highest density intervals (HDIs) around the posterior mean, where $\theta \in$ 95% HDI would indicate a 95% credibility that the true parameter value lies within this range (Kruschke, 2014; McElreath, 2015).

Results

Descriptive statistics and correlations

Descriptive statistics (Table 2) show consistent distributions across E1 and E2. Both participant samples were left-leaning, with 60.7% (E1) and 57.5% (E2) scoring an average political position below four, as opposed to 20.4% (E1) and 19.3% (E2) above four on the seven-point Likert scale, where higher scores indicate greater conservatism. Replicating Schepman and Rodway's (2020, 2022) results with good internal consistency, participants in both samples generally held positive attitudes towards beneficial utilities of AI ($\alpha_{\mathrm{E1}} = 0.89$, $\alpha_{\mathrm{E2}} = 0.88$) and lower dystopian concern ($\alpha_{\mathrm{E1}} = 0.83$, $\alpha_{\mathrm{E2}} = 0.84$). Participants also showed a moderate willingness to accept and act on the statistical AI verdicts of potential prejudice, placed trust in the AI system to detect such anomalies and perceived the AI judgements as fair.

Bayesian Pearson's zero-order correlations for E1 and E2 are displayed in Table 3. In E2 only, participant political position was negatively related to both subscales of general AI attitudes. The latter were correlated with each other in both E1 and E2, as higher scores on both positive and (reverse-coded) negative GAAIS subscales indicated more positive attitude towards AI. In E1 only, both GAAIS subscales weakly correlated only with trust in AI and perceived fairness of AI. Notably, in both experiments, trust and perceived fairness were more strongly correlated to each other than either was to willingness to act.

Bayesian regression analysis

Our pre-registered model simultaneously predicted ratings on all three outcome variables (Table 4 & Figure 1). Ratings on willingness to act were negatively predicted by increasing participant political conservatism in E1 ($\beta = -.16$, 95% HDI = [−0.29, −0.03]) and by the conservative moral intuitive context of AI verdicts in both E1 ($\beta = -.59$, 95% HDI = [−0.77, −0.41]) and E2 ($\beta = -.46$, 95% HDI = [−0.69, −0.24]), suggesting that conservatism of both participants and the context was related to less willingness to act on AI verdicts of potential transgression. In E1, but not E2, the conservative context also predicted less trust in AI ($\beta = -.25$, 95% HDI = [−0.43, −0.07]) and less perceived fairness of AI ($\beta = -.26$, 95% HDI = [−0.43, −0.08]). Additionally, in E1, those who viewed AI in a positive light were more likely to trust and judge the AI as being fair: both positive ($\beta = .16$, 95% HDI = [0.06,

TABLE 2 Descriptive summaries of measured variables in Experiment 1 and 2.

	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	6	3.47 (1.34)	4	6
Social issues	3.14 (1.39)	3	6	3.16 (1.32)	3	6
Foreign policy issues	3.37 (1.34)	4	6	3.40 (1.40)	4	6
Mean political position	3.30 (1.25)	3.33	5.67	3.34 (1.25)	3.33	6
General attitudes towards AI (5-point Likert scale; higher score indicates positive attitudes)						
Positive subscale	3.33 (0.60)	3.33	2.75	3.31 (0.60)	3.33	3.5
Negative subscale	2.97 (0.65)	3	3.25	3.04 (0.69)	3.12	3.75
Responses to scenarios (1 = *Strongly Disagree*, 5 = *Strongly Agree*)						
Willingness to act	3.94 (0.91)	4.07	4	3.90 (0.93)	4.07	4
Trust	3.56 (0.86)	3.63	4	3.44 (0.92)	3.7	4
Perceived fairness	3.68 (0.92)	3.95	4	3.56 (0.94)	3.78	4

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

TABLE 3 Bayesian Pearson's zero-order correlations and their 95% HDIs between main variables in Experiment 1 (E1; lower diagonal) and Experiment 2 (E2; upper diagonal).

	E2					
E1	**Political positions**	**GAAIS positive**	**GAAIS negative**	**Willingness to act**	**Trust**	**Perceived fairness**
Political positions	1	−0.13** [−0.24, −0.02]	−0.13** [−0.25, −0.02]	0.03 [−0.09, 0.14]	0 [−0.1, 0.12]	−0.06 [−0.17, 0.06]
GAAIS positive	−0.07 [−0.16, 0.03]	1	0.5*** [0.41, 0.58]	0.07 [−0.04, 0.18]	−0.01 [−0.12, 0.11]	0.02 [−0.09, 0.13]
GAAIS negative	0.05 [−0.04, 0.14]	0.51*** [0.43, 0.57]	1	0 [−0.11, 0.11]	−0.02 [−0.13, 0.09]	0.02 [−0.09, 0.14]
Willingness to act	0 [−0.1, 0.1]	0.07 [−0.03, 0.17]	0.06 [−0.03, 0.16]	1	0.35*** [0.26, 0.45]	0.36*** [0.26, 0.46]
Trust	−0.02 [−0.12, 0.07]	0.19*** [0.1, 0.29]	0.15*** [0.05, 0.24]	0.3*** [0.21, 0.39]	1	0.63*** [0.56, 0.69]
Perceived fairness	−0.06 [−0.16, 0.03]	0.21*** [0.11, 0.3]	0.11* [0.01, 0.2]	0.36*** [0.27, 0.44]	0.61*** [0.55, 0.67]	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). GAAIS negative values are reverse-coded.

*pd. > 97.5%.

**pd. > 99.5%.

***pd. > 99.95%.

TABLE 4 Summaries of Bayesian regression results in Experiment 1 and 2.

	Willingness to act		Trust		Fairness perception	
	Mean [95% HDI]	*SD*	Mean [95% HDI]	*SD*	Mean [95% HDI]	*SD*
Experiment 1						
Intercept	0.06 [−0.44, 0.56]	0.25	0.16 [−0.31, 0.63]	0.24	0.18 [−0.29, 0.65]	0.24
Political position	**−0.16 [−0.29, −0.03]**	0.07	−0.08 [−0.21, 0.05]	0.07	−0.12 [−0.25, 0.02]	0.07
Context	**−0.59 [−0.77, −0.41]**	0.09	**−0.25 [−0.43, −0.07]**	0.09	**−0.26 [−0.43, −0.08]**	0.09
GAAIS positive	0.08 [−0.01, 0.18]	0.05	**0.16 [0.06, 0.25]**	0.05	**0.18 [0.09, 0.28]**	0.05
GAAIS negative	0.07 [−0.02, 0.16]	0.05	**0.12 [0.02, 0.21]**	0.05	0.07 [−0.03, 0.17]	0.05
Age	0.01 [0, 0.01]	0	0 [−0.01, 0.01]	0	0 [−0.01, 0.01]	0
Political position × Context	**0.31 [0.15, 0.47]**	0.08	0.13 [−0.03, 0.28]	0.08	0.14 [−0.01, 0.28]	0.07
Experiment 2						
Intercept	0.35 [−0.19, 0.89]	0.27	0.08 [−0.44, 0.59]	0.26	0 [−0.51, 0.51]	0.26
Political position	−0.01 [−0.16, 0.15]	0.08	−0.07 [−0.23, 0.09]	0.08	−0.16 [−0.32, 0]	0.08
Context	**−0.46 [−0.69, −0.24]**	0.11	−0.14 [−0.37, 0.09]	0.12	−0.11 [−0.33, 0.13]	0.12
GAAIS positive	0.1 [−0.01, 0.2]	0.05	0.06 [−0.04, 0.16]	0.05	0.07 [−0.03, 0.18]	0.05
GAAIS negative	0.02 [−0.08, 0.13]	0.05	0.05 [−0.05, 0.15]	0.05	0.08 [−0.03, 0.18]	0.05
Age	0 [−0.01, 0.01]	0	0 [−0.01, 0.01]	0	0 [−0.01, 0.01]	0
Political position × Context	**0.19 [0, 0.37]**	0.09	**0.2 [0.02, 0.38]**	0.09	**0.24 [0.06, 0.41]**	0.09

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

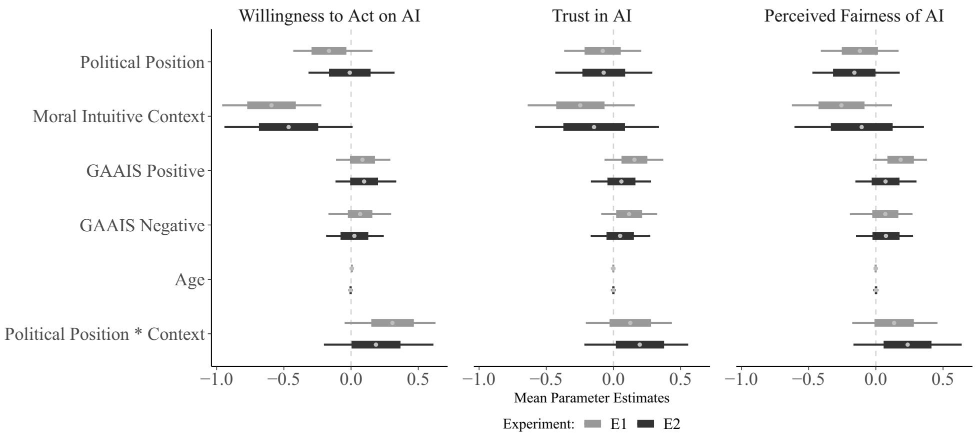

FIGURE 1 Parameter estimates for Willingness to Act on AI verdicts, Trust in AI and Perceived Fairness of AI in E1 & E2, with boxes indicating 95% HDIs and whiskers indicating 100% HDIs. Higher standardized scores on political position correspond to increasing conservatism. Context is a binary variable with liberal/left-wing direction as the reference level.

0.25]) and reversed negative general AI attitudes ($\beta = .12$, 95% HDI = [0.02, 0.21]) predicted more trust in AI, and positive general AI attitudes predicted greater perceived fairness of AI ($\beta = .18$, 95% HDI = [0.09, 0.28]).

We found a belief alignment effect, indexed by the interaction between participant political conservatism and the conservative moral intuitive context of AI verdicts, for all three responses: willingness to act ($\beta = .19$, 95% HDI = [0, 0.37]), trust ($\beta = .2$, 95% HDI = [0.02, 0.38]) and perceived fairness of AI ($\beta = .24$, 95% HDI = [0.06, 0.41]) in E2, though this was evident only for willingness to act ($\beta = .31$, 95% HDI = [0.15, 0.47]) in E1. However, these effects were in the opposite direction of our prediction (Figure 2) – left-wing/liberals showed a stronger increase in Willingness to Act, Trust and Perceived Fairness when the scenario matched their political beliefs, compared to right-wing/conservatives.

We explored the possibility that belief alignment would increase as positive AI attitudes rose, but model comparisons (LOO-IC, ELPD, Bayes factor and Bayesian r^2) indicated such effects did not consistently improve predictive power and strongly favoured the simpler pre-registered models (see Appendix E.4).

Discussion

Two main findings emerge from these experiments. First, participants' judgements of trust, fairness perception, and in particular, willingness to act on AI verdicts of moral transgression were lower in contexts that matched conservative moral intuitions, compared to liberal ones. This may be explained by the conservative contexts violating the politico-moral beliefs of our left-leaning sample of participants. Second, we found direct evidence that belief alignment increased participants' willingness to act, trust and fairness perception, over and above inconsistent effects of general intuitions about AI itself. That is, outcomes were largely driven by an alignment between AI recommendation and pre-existing politico-moral intuitions probed by the scenario contexts, which echoes motivated social cognition needs (Jost, 2017; Jost et al., 2003, 2017; Jost & Amodio, 2012; Jost & Krochik, 2014; Kahan, 2016a, 2016b; Krochik & Jost, 2011; Moore et al., 2021). However, although the effect was evident for all three outcomes in E2, it was only meaningful for willingness to act in E1, possibly because our operationalization of belief alignment did not adequately track the relevance of the deployment contexts (LGBTQ+ rights and environmental protection) to our sample's identities.

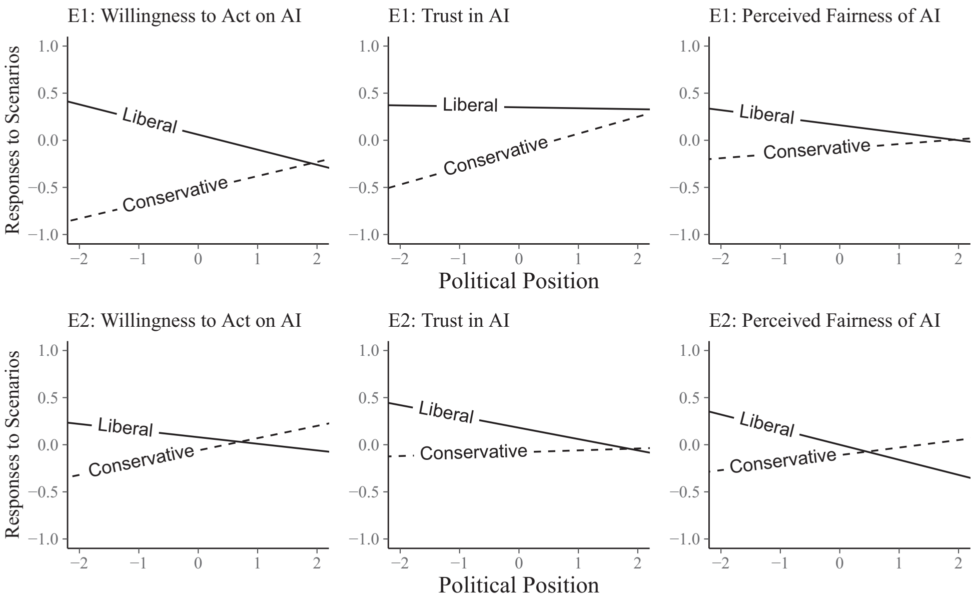

FIGURE 2 Interactions between participant political position and moral intuitive context of AI verdicts for Willingness to Act on AI verdicts, Trust in AI and Perceived Fairness of AI in E1 & E2. Higher scores on political position correspond to increasing conservatism. Solid and dotted lines indicate liberal and conservative contexts of AI deployment, respectively.

While people do have a general political identity that drives affective intuitions (Baldassarri & Page, 2021; Iyengar et al., 2019), they tend to lack ideological coherence (Kalmoe, 2020) and hold moral/political beliefs that diverge from their self-identified partisan stances on some issues (Smith, 2019). As such, even contentious issues are not monolithic on either end of the political spectrum, and the three-item scale for overall political position may be inadequate for capturing issue-specific intuitions probed by the scenarios of AI use. We will address this by using measurements targeted for our topical foci in Experiment 3. Additionally, our measure of fairness did not distinguish between the fairness of the outcome of AI verdicts or involving AI in the process of judgements in these situations. This is akin to the distinction between distributive fairness (Ambrose & Arnaud, 2013) versus procedural fairness (Cropanzano & Ambrose, 2001), which we aim to disentangle in Experiment 3. Lastly, our scenario design attempted to probe a default willingness to initiate an investigation as the direct result of AI verdicts, without explicitly stating it as a consequence of the detected prejudice in the vignettes; we will clarify this in Experiment 3.

EXPERIMENT 3

We conducted E3 to address the limitations of E1 and E2 outlined above, namely the lack of precision in operationalizing belief (mis)alignment, the entanglement of distributive and procedural fairness concerns, and the unclear implication of AI verdicts of moral transgression for the characters in the vignettes. We directly measured participants' endorsement for specific issues on LGBTQ+ rights and environmental protection in addition to overall political orientation, and modified scenarios and probing questions as appropriate.

Since we operationalize belief alignment via new measures here and as we did not find the predicted belief alignment effects for trust in the AI and perceived fairness of the AI in E1, we only predict: (1) increased willingness to act on AI recommendations as a function of belief alignment and (2) belief alignment will remain over and above general AI attitudes.

Methods

Participants

After removing three participants who failed the attention check, the final data set contains responses collected from 302 native English-speaking adults (183 males and 116 females; $M_{age} = 36.83$ years, $SD_{age} = 10.79$ years) with no prior involvement in the study. Testing was conducted on Qualtrics via Prolific, for which each subject received £1.20.

Design, materials and procedures

Experimental set-up for E3 was similar to E1, with modifications to the scenario wording, probe questions and measurements of pre-existing political beliefs. To make explicit the default initiation of investigative actions as the direct consequence of AI detection of potential prejudice, each scenario ended with an extra sentence ('Based on the AI's recommendation, the court/bank has opened an investigation on this judge/loan manager'). Accordingly, willingness to act on AI recommendations became participant's support to uphold the decision to investigate in line with the AI's verdicts ('I think that this judge/loan manager should be suspended until the investigation concludes'). This was intended to emphasize a specific action occurring over a simple agreement that some relatively undefined action should occur. The probe question of trust in AI remained unchanged. To better reflect the psychological complexity of fairness judgements, perceived fairness of AI was split into the perception of procedural fairness of using an AI system for such purposes ('I believe that it's fair to use AI to assess whether a judge/loan manager is biased') and distributive fairness of the AI verdict itself ('I believe that the AI's recommendation to investigate is fair').

Finally, to better capture pre-existing political positions on our specific scenarios, we added four questions each on LGBTQ+ rights and environmental protection, along with filler items on person-centred (people of wealth) and cause-centred (social media) issues as distractions. For each issue, participants responded to a thermometer scale ('On a scale of 0–100, how warm/cold do you feel about X?'), the extent of concern ('How concerned are you about X?'; 1 = *not at all concerned* and 5 = *very concerned*) and two more questions on relevant personal experiences (data collected but not analyzed due to a technical error, and pre-registration was updated prior to analysis; see Appendix B).

Statistical analysis plan

Our analyses here are largely identical to E1, with the modification of having four simultaneous dependent variables (willingness to act, trust in AI, perceived procedural fairness, perceived distributive fairness) predicted by fixed effects of context, issue-specific attitudes and the interaction of the two terms. Participant political position was kept as a covariate to control for its influence on the dependent variables and correlations with issue-specific attitudes. Other parameters included covariates of GAAIS positive and negative subscale means, nuisance variable of age, as well as random effects of scenario, topic and individual subject. We standardized all variables as appropriate. To obtain continuous measures of issue-specific attitudes, we averaged the standardized thermometer scale responses and extent of concern of the two topical foci, with higher scores indicating more endorsement for LGBTQ+ rights or environmental protection.

Using the *brms* package (v. 2.19.0; Bürkner, 2017, 2018) in RStudio (v. 4.3.0; R Core Team, 2023), we estimated Bayesian multilevel multivariate multiple regression models containing parameters specified above to predict all four outcomes for each focus separately. Additionally, we explored using overall political position and its interaction with context as main predictors without issue-specific endorsement to compare with general trends observed in E1. We used the posterior means and *SDs* from E1 (Table 4, top panel) as priors in E3. In cases of absent priors, that is for issue-specific endorsement and its interaction with context, we used posterior estimates of overall political position and its interaction terms from E1, with signs reversed to indicate greater endorsement for LGBTQ+ rights and environmental protection as the opposite of increasing conservatism.

Results

Descriptive statistics and correlations

Table 5 displays descriptive statistics for E3. Similar to E1 and E2, this sample was largely left-leaning, with 58.6% scoring an average below four and 13.6% above four on a seven-point Likert scale across economic, social and foreign policy issues. The overall liberal political orientation was additionally reflected in the reporting of warm feelings and high levels of concern towards LGBTQ+ rights and environmental protection. Closely replicating Schepman and Rodway's (2020, 2022) results once more with both subscales showing good internal consistency ($\alpha_{\text{PosAtt}} = 0.9$, $\alpha_{\text{NegAtt}} = 0.83$), participants reported positive general attitudes towards practical benefits of AI and held less dystopian concerns over AI. Finally, scenario responses revealed a moderate willingness to uphold investigative actions following AI detection of potential prejudice, and that on average, subjects trusted the AI's recommendation, considered it fair to use AI for such purposes and perceived AI verdict outcomes as fair.

Table 6 below shows a somewhat different correlational pattern relative to the previous two experiments. First, unsurprisingly, political conservatism showed a moderate negative correlation with endorsement for LGBTQ+ rights and environmental protection, which were moderately positively correlated themselves. Second, positive and reverse-coded negative GAAIS subscales were positively

TABLE 5 Descriptive summaries of measured variables in Experiments 3.

	Mean (*SD*)	Median	Range
Political positions (1 = *Very Left/Liberal*, 7 = *Very Right/Conservative*)			
Economic issues	3.25 (1.38)	3	6
Social issues	3.05 (1.35)	3	6
Foreign policy issues	3.30 (1.35)	4	6
Mean political position	3.20 (1.28)	3.33	6
Issue-specific attitudes (higher score indicates positive attitudes)			
LGBTQ+ rights			
Temperate scale (1° ~ 100°)	72.22 (29.97)	82	100
Extent of concern (5-point Likert scale)	3.40 (1.16)	4	4
Environmental protection			
Temperate scale (1° ~ 100°)	73.55 (22.18)	78	100
Extent of concern (5-point Likert scale)	4.06 (0.88)	4	4
General attitudes towards AI (5-point Likert scale; higher score indicates positive attitudes)			
Positive subscale	3.23 (0.63)	3.33	3.75
Negative subscale	3.03 (0.68)	3.12	3.5
Responses to scenarios (1 = *Strongly Disagree*, 5 = *Strongly Agree*)			
Willingness to act	3.15 (1.16)	3.04	4
Trust	3.39 (0.94)	3.49	4
Procedural fairness	3.74 (0.92)	3.99	4
Distributive fairness	3.36 (1.05)	3.53	4

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

TABLE 6 Bayesian Pearson's zero-order correlations and their 95% HDIs between main variables in Experiment 3.

	Political positions	LGBTQ+ rights	Environmental protection	GAAIS positive	GAAIS negative	Willingness to act	Trust	Procedural fairness
LGBTQ+ rights	−0.46*** [−0.53, −0.41]							
Environmental protection	−0.34*** [−0.41, −0.26]	0.49*** [0.43, 0.55]						
GAAIS positive	−0.12** [−0.19, −0.04]	0.22*** [0.14, 0.29]	0.22*** [0.15, 0.30]					
GAAIS negative	0.08* [0, 0.15]	0.08* [0, 0.16]	−0.04 [−0.12, 0.04]	0.36*** [0.29, 0.43]				
Willingness to act	−0.12*** [−0.20, −0.04]	0.19*** [0.12, 0.27]	0.17*** [0.09, 0.24]	0.11** [0.03, 0.18]	0.04 [−0.04, 0.12]			
Trust	−0.04 [−0.12, 0.04]	0.16*** [0.08, 0.24]	0.27*** [0.20, 0.34]	0.41*** [0.34, 0.47]	0.25*** [0.17, 0.32]	0.46*** [0.39, 0.52]		
Procedural fairness	−0.06 [−0.14, 0.02]	0.18*** [0.11, 0.26]	0.25*** [0.17, 0.32]	0.28*** [0.21, 0.36]	0.16*** [0.07, 0.23]	0.48*** [0.42, 0.54]	0.61*** [0.56, 0.66]	
Distributive fairness	−0.01 [−0.09, 0.07]	0.17*** [0.09, 0.24]	0.25*** [0.18, 0.33]	0.39*** [0.32, 0.45]	0.23*** [0.16, 0.31]	0.35*** [0.28, 0.42]	0.58*** [0.53, 0.63]	0.55*** [0.49, 0.61]

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). GAAIS negative values are reverse-coded.

*pd. > 97.5%.

**pd. > 99.5%.

***pd. > 99.95%.

correlated; while the former was linked to all four outcomes, greater support for LGBTQ+ rights and environmental protection, and less overall conservatism, the latter was linked to most outcomes except willingness to act, and weakly associated with overall political conservatism and endorsement for LGBTQ+ rights. Lastly, all four outcomes showed a positive relationship amongst themselves and with issue-specific endorsement, although only willingness to act was negatively related to overall political conservatism.

Pre-registered analysis

Next, we conducted pre-registered analyses where issue-specific attitudes towards LGBTQ+ rights (Table 7a) and environmental protection (Table 7b) were modelled separately while controlling for general political orientation, which produced strikingly similar results (Figure 3). All four outcomes meaningfully increased as a function of increasing individual support for both issues ($.11 \leq$ all βs $\leq .21$), but scored lower in the conservative context ($-.51 \leq$ all βs $\leq -.19$). Greater positivity towards AI's usefulness predicted greater trust and perception of procedural/distributive fairness across both issues ($.17 \leq$ all βs $\leq .26$); lower dystopian AI concerns showed a similar pattern ($.09 \leq$ all βs $\leq .15$) except for trust in the environmentalism condition. Ratings for willingness to act stood out, only decreasing with greater participant conservatism ($\beta_a = \beta_b = -.15$).

Finally, belief (mis)alignment effect was found for both issues, not only for willingness to act as hypothesised ($\beta_a = -.22$, 95% HDI = [−0.34, −0.1]; $\beta_b = -.29$, 95% HDI = [−0.41, −0.17]), but also for procedural fairness ($\beta_a = -.18$, 95% HDI = [−0.29, −0.07]; $\beta_b = -.16$, 95% HDI = [−0.26, −0.05]) and distributive fairness ($\beta_a = -.13$, 95% HDI = [−0.24, −0.02]; $\beta_b = -.12$, 95% HDI = [−0.22, −0.01]), and for trust but only in the LGBTQ+ rights condition ($\beta_a = -.12$, 95% HDI = [−0.24, 0]). That is, issue-specific attitudes and context interacted, such that with conservative AI verdicts, those with greater concern for LGBTQ+ rights and environmental protection were less willing to act on such verdicts, considered involving AI in the decision procedure and the outcome less fair and trusted the AI less.

Exploratory analysis

Additional to the pre-registered analyses, we replicated E1 analysis estimating belief alignment with overall political position without issue-specific endorsement, returning similar patterns of results: ratings for outcome variables reduce with participant and context conservatism and increase with greater positivity towards AI. Unlike E1, however, belief alignment was meaningful for all outcomes ($.1 \leq$ all βs $\leq .21$; see Appendix E.1). We also explored three-way interactions between context, issue-specific/overall political positions and positive/negative general AI attitudes; as in E1 & E2, the more complex model specification did not improve predictive power, again strongly favouring simpler models (see Appendix E.4).

Discussion

E3 generated three main results. First, as in E1 and E2, participants' willingness to act on AI verdicts and judgements of trust and fairness was higher in the liberal context and stronger as a function of increasing support for LGBTQ+ rights and environmental protection; unsurprising, given this sample was, again, left-leaning and reported strong endorsement for both issues.

Second, positivity towards AI predicted greater trust and perceptions of both procedural and distributive fairness. This may stem from the presentation of AI verdicts as purely statistical patterns, giving the impression that prejudice, a morally charged phenomenon reduced to statistical anomalies, is measurable and quantifiable. In line with previous findings of task-specific algorithmic aversion towards delegating

TABLE 7 Summaries of Bayesian regression results in Experiment 3 (pre-registered models).

	Willingness to act		Trust		Procedural fairness		Distributive fairness	
	Mean [95% HDI]	***SD***	**Mean [95% HDI]**	***SD***	**Mean [95% HDI]**	***SD***	**Mean [95% HDI]**	***SD***
(a) LGBTQ+ rights								
Intercept	0.2 [−0.26, 0.64]	0.23	0.13 [−0.21, 0.54]	0.18	0.3 [−0.04, 0.7]	0.18	0.11 [−0.23, 0.52]	0.18
LGBTQ+ attitudes	**0.21** **[0.11, 0.31]**	0.05	**0.11** **[0.02, 0.21]**	0.05	**0.18** **[0.08, 0.27]**	0.05	**0.11** **[0.01, 0.2]**	0.05
Context	**−0.5** **[−0.63, −0.37]**	0.07	**−0.22** **[−0.35, −0.1]**	0.07	**−0.29** **[−0.42, −0.17]**	0.06	**−0.21** **[−0.34, −0.09]**	0.06
GAAIS positive	0.04 [−0.03, 0.11]	0.04	**0.26** **[0.19, 0.33]**	0.04	**0.16** **[0.09, 0.23]**	0.04	**0.25** **[0.18, 0.32]**	0.04
GAAIS negative	0.05 [−0.02, 0.12]	0.04	**0.12** **[0.05, 0.18]**	0.04	**0.09** **[0.02, 0.16]**	0.04	**0.1** **[0.04, 0.17]**	0.03
Political position	**−0.15** **[−0.24, −0.07]**	0.04	−0.04 [−0.13, 0.04]	0.04	−0.08 [−0.17, 0.01]	0.04	−0.07 [−0.15, 0.02]	0.04
Age	0 [0, 0.01]	0	0 [0, 0.01]	0	0 [−0.01, 0]	0	0 [0, 0.01]	0
LGBTQ+ attitudes × Context	**−0.22** **[−0.34, −0.1]**	0.06	**−0.12** **[−0.24, 0]**	0.06	**−0.18** **[−0.29, −0.07]**	0.06	**−0.13** **[−0.24, −0.02]**	0.06
(b) Environmental concerns								
Intercept	0.06 [−0.3, 0.5]	0.2	0.18 [−0.15, 0.58]	0.18	0.23 [−0.15, 0.68]	0.21	0.06 [−0.29, 0.5]	0.2
Environmentalist attitudes	**0.15** **[0.05, 0.24]**	0.05	**0.12** **[0.04, 0.21]**	0.04	**0.19** **[0.1, 0.28]**	0.05	**0.19** **[0.1, 0.28]**	0.05
Context	**−0.51** **[−0.64, −0.38]**	0.07	**−0.19** **[−0.3, −0.07]**	0.06	**−0.34** **[−0.47, −0.22]**	0.06	**−0.27** **[−0.4, −0.15]**	0.06
GAAIS positive	0.05 [−0.02, 0.13]	0.04	**0.23** **[0.17, 0.3]**	0.03	**0.17** **[0.1, 0.24]**	0.04	**0.24** **[0.17, 0.31]**	0.04
GAAIS negative	0.04 [−0.04, 0.11]	0.04	**0.15** **[0.09, 0.22]**	0.03	0.06 [0, 0.13]	0.03	**0.1** **[0.04, 0.17]**	0.04

TABLE 7 (Continued)

	Willingness to act		Trust		Procedural fairness		Distributive fairness	
	Mean [95% HDI]	***SD***	**Mean [95% HDI]**	***SD***	**Mean [95% HDI]**	***SD***	**Mean [95% HDI]**	***SD***
Political position	**-0.15** **[−0.24, −0.06]**	0.04	-0.06 [−0.14, 0.02]	0.04	−0.07 [−0.16, 0.01]	0.04	−0.01 [−0.1, 0.07]	0.04
Age	0 [0, 0.01]	0	0 [−0.01, 0]	0	0 [−0.01, 0.01]	0	0 [0, 0.01]	0
Env. attitudes × Context	**−0.29** **[−0.41, −0.17]**	0.06	−0.04[−0.14, 0.07]	0.06	**−0.16** **[−0.26, −0.05]**	0.05	**−0.12** **[−0.22, −0.01]**	0.06

Note: Model converged successfully with split $\hat{R} = 1$ for all estimated parameters. Context is a binary variable with liberal/left-wing direction as the reference level. GAAIS negative values are reverse-coded. Bold emphasizes $0 \notin$ 95% HDI.

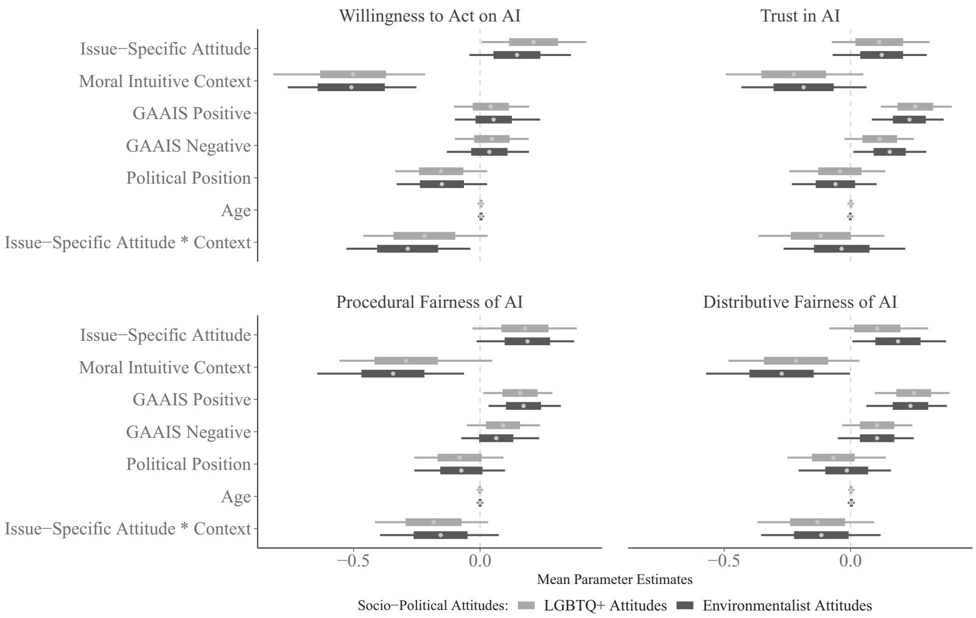

FIGURE 3 Parameter estimates for Willingness to Act based on AI verdicts, Trust in AI and perception of Procedural Fairness and Distributive Fairness of AI in E3, with boxes indicating 95% HDIs and whiskers indicating 100% HDIs. Higher standardized scores on political position correspond to increasing conservatism. Context is a binary variable with liberal/left-wing direction as the reference level.

to AI perceived subjective tasks (Castelo et al., 2019; Lee, 2018), our results suggest that the perception of AI use in moral situations commonly considered to require unique human abilities, such as freewill and autonomy (Jauernig et al., 2022), may be improved by the perceived objectivity of the task at hand.

Third and most importantly, we reproduced belief alignment for both overall political position and specific endorsement for LGBTQ+ rights and environmental protection (only except trust in the environmentalism condition). This suggests judgements towards AI verdicts of moral transgression were in part driven by the alignment of the verdict with participants' pre-existing politico-moral beliefs in those contexts, which are themselves correlated with overall political leaning. The consistent pattern of belief alignment here across overall political orientation and specific endorsement eased our previous concern that general political position might be inadequate at capturing issue-specific intuitions probed by our chosen contexts. Furthermore, general AI attitudes were as strong as, if not stronger than, belief alignment itself for trust and fairness judgements, but not for willingness to act. This suggests a possible tension that despite judging the AI and its output as trustworthy and/or fair, participants were not always willing to act on its verdicts. Aydin and Malle (2024) report a similar disassociation between processes of deliberating the content of the advice and assessing the characteristics of advisors (e.g. trustworthiness). They found that while the persuasion effect of both AI and human legal advisors was equally strong, human advisors received greater approval and trust, compared to AI advisors. These findings highlight the nuances involved in AI-mediated moral judgements that require further research.

GENERAL DISCUSSIONS

Summary of main findings

Using morally charged vignettes, we investigated whether participants' judgements towards AI decisions are driven by their belief alignment with the underlying context of AI deployment over and above general AI attitudes. Across three experiments, we found evidence that (1) while positivity about AI influenced judgements towards its deployment (especially for judgements of trust and fairness), they were also uniquely driven by the compatibility between AI verdicts and participants' existing moral/political values beyond general attitudes towards AI itself; (2) these effects were often in tension with one another and of similar magnitude; and that (3) liberal-leaning orientations and contexts predicted more willingness to act on AI verdicts, more trust and greater perception of (procedural/distributive) fairness. These results offer several theoretical implications for judgement and decision-making and practical implications for the future adoption of automated decision-making in morally significant contexts.

Implications, limitations and further research

First, the finding of belief alignment is consistent with research on motivated social cognition where individuals have the tendency to selectively accept/process belief-consistent information or intuitively reject belief-inconsistent information (e.g. Lewandowsky & Oberauer, 2016; Moore et al., 2021; Tucker et al., 2018). Since much research has been conducted on politicized social/scientific beliefs (e.g. Drummond & Fischhoff, 2017; Glinitzer et al., 2021), we provide an example of motivated reasoning in human–computer interaction. It is noteworthy that belief alignment effects are not necessarily indicative of bias and can at times reflect rational information processing principles (Cook & Lewandowsky, 2016). Furthermore, both theoretical simulations (Hahn et al., 2020) and behavioural experiments (Fränken et al., 2020, 2024) suggest that inference on social information does not account for the structure of information dependency – people do not appear to distinguish hearsay from evidence, and trust may play a significant role in moderating the force of such socially driven learning. Thus, our results that positive AI attitudes predict increased trust in, perceived fairness of, and occasionally willingness to act on AI verdicts of human transgression can act as a counterweight to belief misalignment.

The increase in trust and fairness judgements, driven by general optimism towards AI, may pose risks of misuse by corporations or governments seeking to further their own interests (Liu et al., 2022). This is further complicated by the fact that the consultation, delegation or deference of moral judgements to AI can introduce complications in assigning responsibility and blame/praise for multi-agent decision-making (Matthias, 2004; Vallor & Vierkant, 2024). Further research should investigate how individuals assign responsibility, blame and credit when they follow or disregard AI suggestions as a result of belief (mis) alignment. Initial research on the attribution of responsibility involving AI advisors has already shown that compatibility between AI suggestions/advice and one's (meta) desires/intentions could licence a particular decision/behaviour, which might in turn lead to the transfer of responsibility to the enabling agent (Dong & Bocian, 2024). Amidst the increasingly ubiquitous applications of AI, this is particularly crucial for technologies with the potential to profoundly influence many lives, such as medical diagnosis, transplant allocation, loan outcomes and admissions or employment decisions.

Moreover, trust and fairness perception, while positively correlated with attitudes towards the practical benefits of AI, are not reducible to willingness to uphold AI decisions. Fairness perceptions of algorithmic decision outcomes and processes are conceptually linked and have been studied extensively: algorithmic factors such as accuracy, transparency and input features all influence perceived fairness, along with human factors like socio-demographics, self-interest and familiarity (Starke et al., 2022). Trust itself can be separated into functionality-based and human-like trust in AI (Choung et al., 2023), with the former potentially linking to competence and reliability (ability-based performance trust) and the latter to sincerity, integrity and benevolence (human-like moral trust) (Malle & Ullman, 2021; see also Ullman & Malle, 2018, 2019). This aligns with previous findings highlighting an objective–subjective distinction in task characteristics, where tasks with more quantifiable aspects tend to elicit higher levels of trust (Castelo et al., 2019; Lee, 2018). Future studies could clarify the relationships between these different facets of trust across different types of AI applications. However, both fields of study reveal inconclusive results concerning the factors influencing trust(worthiness) and the perception of AI, as well as human–AI comparative effects. It is thus crucial to recognize context dependency in trust and fairness in algorithmic decisions such that a general conclusion across board is likely infeasible.

We further found that despite the prevalent aversion for involving AI in moral decision-making, participants exhibited positive attitudes towards this particular form of algorithmic detection of prejudice. A possible explanation for this could be the reduction of moral transgression (e.g. prejudice) to a simple statistical pattern of favouring/discriminating against particular groups of individuals (e.g. loan application success rate for LGBTQ+ couples). This presentation of prejudice identification as an objective and/or quantifiable task may have somewhat mitigated algorithmic aversion in the moral domain. Considering the inherent morally relevant features in many contexts of AI deployment, however, while this may enhance the acceptability of algorithms, overly reductionistic computerisation of intricate moral matters could lead to long-term harm if fundamental structural issues underpinning, for example, prejudice, remain neglected. This concern is heightened by the rise of popular generative AI technologies, particularly following the release of ChatGPT in 2022: given the ease with which these systems can generate and disseminate false or misleading information, presenting AI-generated content as objective 'ground truth' is potentially dangerous and raises significant ethical challenges.

Methodologically, our characterization of a narrow statistical pattern-detection AI is both a limitation and a strength. It is constrained in scope and somewhat outdated, particularly in comparison to more recent advancements, such as large language models (LLMs) that involve more sophisticated deep learning techniques, trained on vast amounts of data and refined through reinforcement learning from human feedback. A recent large-scale study ($N=4836$) has already shown that LLM-generated messages on political policies are just as persuasive as human-written ones, even for highly polarized topics such as gun control (Bai et al., 2023). As such, further research is needed to explore the impact of belief alignment on more human-like, assertive language coming from AI, given the well-documented risks of such models generating misinformation and nonsensical content (Hicks et al., 2024).

Yet, our description of a rigid, non-agentic AI allows us to isolate the role of belief alignment motivated by underlying contexts of AI deployment. Finding this effect as strong as general AI attitudes towards a simple, static AI speaks to how much more pronounced this effect might be with more complex systems and more information available regarding how they are built, trained, by whom and for what purposes. Given the known and widely reported algorithmic bias (Gebru, 2020; Mehrabi et al., 2021; Weidinger et al., 2022) and growing public concerns towards big tech companies (Ibrahim et al., 2024), such assumptions may amplify the motivated cognition aspects of judgements towards AI output, potentially enhancing the magnitude of belief alignment, especially through feeding back into the training data (Kidd & Birhane, 2023).

Finally, those more familiar or experienced with algorithmic/AI justice literature or technical aspects of AI may interpret our stimuli differently from the general public; we also cannot be certain if participants made other assumptions unaccounted for by our materials. As a result, GAAIS may be insufficient for capturing more nuanced attitudes towards AI, as it may not effectively differentiate between more subtle perspectives or various types of AI applications. While it is adequate for our purposes of gauging general sentiments towards AI amongst the broader public, future research could explore this issue further, potentially adopting a mixed-methods approach that incorporates qualitative data.

CONCLUSIONS

In this three-part pre-registered study, we examined how the alignment of prior socio-moral beliefs with contextual information impacted judgement on AI moral recommendations. Results revealed that both belief alignment and general positivity towards AI influenced willingness to act on AI verdicts, trust in AI and perceived procedural/distributive fairness of AI. As such, previous research on motivated reasoning extends to artificial agents, highlighting the complexity of AI-mediated moral judgements given the emergence of AI as a category of pseudo-moral agents and potential risks in the large-scale deployment of autonomous decision systems, even with human oversight.

AUTHOR CONTRIBUTIONS

Yuxin Liu: Conceptualization; methodology; software; data curation; investigation; validation; formal analysis; funding acquisition; visualization; project administration; resources; writing – original draft; writing – review and editing. **Adam Moore:** Conceptualization; formal analysis; methodology; investigation; software; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

The corresponding author's PhD programme was funded by the Baillie Gifford studentship in AI/Data Ethics at the Centre for Technomoral Futures, Edinburgh Futures Institute. Participants were compensated using the postgraduate research student grant from the School of Philosophy, Psychology and Language Sciences (PPLS), University of Edinburgh. All experiments were approved by the Research Ethics Committee of the School of PPLS (ref: 245-2021/1; 245-2021/2) at the University of Edinburgh.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available on Open Science Framework at https://osf.io/7qjt3/.

ORCID

Yuxin Liu https://orcid.org/0000-0001-9034-0030

REFERENCES

Ambrose, M. L., & Arnaud, A. (2013). Are procedural justice and distributive justice conceptually distinct? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 59–84). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203774847-10

Amodio, D. M., Jost, J. T., Master, S. L., & Yee, C. M. (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience*, *10*(10), 1246–1247. https://doi.org/10.1038/nn1979

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Aydin, Z., & Malle, B. F. (2024). *Dissociated responses to AI: Persuasive but not trustworthy?* Proceedings of the Annual Meeting of the Cognitive Science Society, 46. https://escholarship.org/uc/item/90b426g8

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human-Computer Interaction*, *40*(3), 1–16. https://doi.org/10.1080/10447318.2022.2138826

Bai (Max), H., Voelkel, J. G., Eichstaedt, j. C., & Willer, R. (2023). *Artificial intelligence can persuade humans on political issues*. OSF Preprints https://doi.org/10.31219/osf.io/stakv

Baldassarri, D., & Page, S. E. (2021). The emergence and perils of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), e2116863118. https://doi.org/10.1073/pnas.2116863118

Banks, J. (2020). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, *13*, 2021–2038. https://doi.org/10.1007/s12369-020-00692-3

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science*, *14*(2), 292–303. https://doi.org/10.1177/1745691618788876

Bartels, L. M. (2002). Beyond the running tally: Partisan bias in political perceptions. *Political Behavior*, *24*(2), 117–150. https://doi.org/10.1023/A:1021226224601

Betancourt, M. (2018). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv:1701.02434 [Stat]. http://arxiv.org/abs/1701.02434

Bianchi, E. C., Brockner, J., van den Bos, K., Seifert, M., Moon, H., van Dijke, M., & De Cremer, D. (2015). Trust in decision-making authorities dictates the form of the interactive relationship between outcome fairness and procedural fairness. *Personality and Social Psychology Bulletin*, *41*(1), 19–34. https://doi.org/10.1177/0146167214556237

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology*, *75*(1), 653–675. https://doi.org/10.1146/annurev-psych-030123-113559

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Cram, L., Moore, A., Olivieri, V., & Suessenbach, F. (2018). Fair is fair, or is it? Territorial identity triggers influence ultimatum game behavior. *Political Psychology*, *39*(6), 1233–1250. https://doi.org/10.1111/pops.12543

Cropanzano, R., & Ambrose, M. L. (2001). Procedural and distributive justice are more similar than you think: A monistic perspective and a research agenda. In J. Greenberg & R. Cropanzano (Eds.), *Advances in organizational justice* (pp. 119–150). Stanford University Press.

Crowson, H. M. (2009). Are all conservatives alike? A study of the psychological correlates of cultural and economic conservatism. *The Journal of Psychology*, *143*(5), 449–463. https://doi.org/10.3200/JRL.143.5.449-463

de Cremer, D., & Tyler, T. R. (2007). The effects of trust in authority and procedural fairness on cooperation. *Journal of Applied Psychology*, *92*(3), 639–649. https://doi.org/10.1037/0021-9010.92.3.639

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dong, M., & Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. *Journal of Experimental Social Psychology*, *111*, 104584. https://doi.org/10.1016/j.jesp.2023.104584

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(36), 9587–9592. https://doi.org/10.1073/pnas.1704882114

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127–150. https://doi.org/10.1111/pops.12394

Fränken, J.-P., Theodoropoulos, N. C., Moore, A. B., & Bramley, N. R. (2020). *Belief revision in a micro-social network: Modeling sensitivity to statistical dependencies in social learning*. Proceedings of the 42nd annual meeting of the cognitive science Society, 42, 1255–1261. https://www.bramleylab.ppls.ed.ac.uk/pdfs/fraenken2020belief.pdf

Fränken, J.-P., Valentin, S., Lucas, C. G., & Bramley, N. R. (2024). Naïve information aggregation in human social learning. *Cognition*, *242*, 105633. https://doi.org/10.1016/j.cognition.2023.105633

Gaines, B. J., Kuklinski, J. H., Quirk, P. J., Peyton, B., & Verkuilen, J. (2007). Same facts, different interpretations: Partisan motivation and opinion on Iraq. *The Journal of Politics*, *69*(4), 957–974. https://doi.org/10.1111/j.1468-2508.2007.00601.x

Gebru, T. (2020). Race and gender. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 251–269). Oxford University Press. https://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397-e-16

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press/Taylor & Francis Group. http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=1438153

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, *58*(1), 129–149. https://doi.org/10.1111/bjso.12286

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy and Technology*, *31*(2), 169–188. https://doi.org/10.1007/s13347-017-0285-z

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Glinitzer, K., Gummer, T., & Wagner, M. (2021). Learning facts about migration: Politically motivated learning of polarizing information about refugees. *Political Psychology*, *42*(6), 1053–1069. https://doi.org/10.1111/pops.12734

Grgić-Hlača, N., Lima, G., Weller, A., & Redmiles, E. M. (2022). *Dimensions of diversity in human perceptions of algorithmic fairness*. Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–12. https://doi.org/10.1145/3551624.3555306

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). *Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction*. Proceedings of the 2018 world wide web conference, 903–912 https://doi.org/10.1145/3178876.3186138

Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, *197*(4), 1511–1541. https://doi.org/10.1007/s11229-018-01936-6

Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Harnish, R. J., Bridges, K. R., & Gump, J. T. (2018). Predicting economic, social, and foreign policy conservatism: The role of right-wing authoritarianism, social dominance orientation, moral foundations orientation, and religious fundamentalism. *Current Psychology*, *37*(3), 668–679. https://doi.org/10.1007/s12144-016-9552-x

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hobolt, S. B., Leeper, T. J., & Tilley, J. (2021). Divided by the vote: Affective polarization in the wake of the Brexit referendum. *British Journal of Political Science*, *51*(4), 1476–1493. https://doi.org/10.1017/S0007123420000125

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human Computer Interaction*, *36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Ibrahim, H., Debicki, M., Rahwan, T., & Zaki, Y. (2024). Big tech dominance despite global mistrust. *IEEE Transactions on Computational Social Systems*, *11*(3), 3741–3752. https://doi.org/10.1109/TCSS.2023.3339183

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. *Philosophy and Technology*, *35*(1), 2. https://doi.org/10.1007/s13347-021-00495-y

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, *38*(2), 167–208. https://doi.org/10.1111/pops.12407

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion*, *36*(1), 55–64. https://doi.org/10.1007/s11031-011-9260-7

Jost, J. T., & Krochik, M. (2014). Ideological differences in epistemic motivation: Implications for attitude structure, depth of information processing, susceptibility to persuasion, and stereotyping. In A. J. Elliot (Ed.), *Advances in motivation science* Vol. 1 (pp. 181–231). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/S2215091914000066

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. https://doi.org/10.1037/0033-2909.129.3.339

Jost, J. T., Stern, C., Rule, N. O., & Sterling, J. (2017). The politics of fear: Is there an ideological asymmetry in existential motivation? *Social Cognition*, *35*(4), 324–353. https://doi.org/10.1521/soco.2017.35.4.324

Kahan, D. M. (2016a). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In R. A. Scott, S. M. Kosslyn, & M. C. Buchmann (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (1st ed.). Wiley. https://doi.org/10.1002/9781118900772.etrds0417

Kahan, D. M. (2016b). The politically motivated reasoning paradigm, part 2: Unanswered questions. In R. A. Scott, S. M. Kosslyn, & M. C. Buchmann (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (1st ed.). Wiley. https://doi.org/10.1002/9781118900772.etrds0418

Kahn, P. H., Severson, R. L., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., & Freier, N. G. (2012). *Do people hold a humanoid robot morally accountable for the harm it causes?* Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, 33–40. https://doi.org/10.1145/2157689.2157696

Kalmoe, N. P. (2020). Uses and abuses of ideology in political psychology. *Political Psychology*, *41*(4), 771–793. https://doi.org/10.1111/pops.12650

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, *380*(6651), 1222–1223. https://doi.org/10.1126/science.adi0248

Krochik, M., & Jost, J. T. (2011). Ideological conflict and polarization: A social psychological perspective. In D. Bar-Tal (Ed.), *Intergroup conflicts and their resolution: A social psychological perspective* (pp. 145–174). Psychology Press. https://www.taylorfrancis.com/chapters/mono/10.4324/9780203834091-13

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press. https://www-sciencedirect-com/book/9780124058880/doing-bayesian-data-analysis

Ladak, A., Loughnan, S., & Wilks, M. (2023). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*, 27–34. https://doi.org/10.1177/09637214231205866

Lara, F., & Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*, *13*(3), 275–287. https://doi.org/10.1007/s12152-019-09401-y

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017). *Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division*. Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 1035–1048. https://doi.org/10.1145/2998181.2998230

Lee, M. K., & Rich, K. (2021). *Who is included in human perceptions of AI? Trust and perceived fairness around healthcare AI and cultural mistrust*. Proceedings of the 2021 CHI conference on human factors in computing systems, 1–14. https://doi.org/10.1145/3411764.3445570

Lewandowsky, S., & Oberauer, K. (2016). Motivated rejection of science. *Current Directions in Psychological Science*, *25*(4), 217–222. https://doi.org/10.1177/0963721416654436

Liang, G., & Newell, B. (2022). *Trusting algorithms: Performance, explanations, and sticky preferences*. Proceedings of the annual meeting of the cognitive science Society, 44. https://escholarship.org/uc/item/64x316kg.

Liu, Y., Moore, A., Webb, J., & Vallor, S. (2022). *Artificial moral advisors: A new perspective from moral psychology*. Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and Society, 436–445. https://doi.org/10.1145/3514094.3534139

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Academic Press. https://www.sciencedirect.com/science/article/pii/B9780128194720000010

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. S. Sequeira, G. S. Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* Vol 95 (pp. 111–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_1110.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). *Sacrifice one for the good of many?: People apply different moral norms to human and robot agents*. Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, 117–124 https://doi.org/10.1145/2696454.2696458

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly*, *52*(2), 155–194. https://doi.org/10.1111/j.1533-8525.2011.01198.x

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan* (1st ed.). CRC Press/Taylor & Francis Group. https://ebookcentral.proquest.com/lib/ed/detail.action?docID=4648054

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Moore, A., Hong, S., & Cram, L. (2021). Trust in information, political identity and the brain: An interdisciplinary fMRI study. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *376*(1822), 20200140. https://doi.org/10.1098/rstb.2020.0140

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*, 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Prahl, A., & van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702. https://doi.org/10.1002/for.2464

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65. https://doi.org/10.1057/s41599-019-0279-9

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports*, *1*, 100014. https://doi.org/10.1016/j.chbr.2020.100014

Schepman, A., & Rodway, P. (2022). The general attitudes towards artificial intelligence scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction*, *39*(8), 1–18. https://doi.org/10.1080/10447318.2022.2085400

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shank, D. B., Bowen, M., Burns, A., & Dew, M. (2021). Humans are perceived as better, but weaker, than artificial intelligence: A comparison of affective impressions of humans, AIs, and computer systems in roles on teams. *Computers in Human Behavior Reports*, *3*, 100092. https://doi.org/10.1016/j.chbr.2021.100092

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6

Skitka, L. J., & Mullen, E. (2002). Understanding judgments of fairness in a real-world political context: A test of the value protection model of justice reasoning. *Personality and Social Psychology Bulletin*, *28*(10), 1419–1429. https://doi.org/10.1177/014616702236873

Smith, M. (2019). *Left-wing vs right-wing: It's complicated*. YouGov. https://yougov.co.uk/topics/politics/articles-reports/2019/08/14/left-wing-vs-right-wing-its-complicated

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 20539517221115189. https://doi.org/10.1177/20539517221115189

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139

Tyler, T. R., & Degoey, P. (1996). Trust in organizational authorities: The influence of motive attributions on willingness to accept decisions. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 331–356). Sage Publications. http://knowledge.sagepub.com/view/trust-in-organizations/SAGE.xml

Tyler, T. R., & Smith, H. J. (1999). Justice, social identity, and group processes. In T. R. Tyler, R. M. Kramer, & O. P. John (Eds.), *The psychology of the social self* (1st ed., pp. 223–264). Psychology Press. https://www.taylorfrancis.com/books/9781317778288

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022). *Trust in human-AI interaction: Scoping out models, measures, and methods*. Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3491101.3519772

Ullman, D., & Malle, B. F. (2018). *What does it mean to trust a robot? Steps toward a multidimensional measure of trust*. Companion of the 2018 ACM/IEEE international conference on human-robot interaction, 263–264. https://doi.org/10.1145/3173386.3176991

Ullman, D., & Malle, B. F. (2019). *Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust*. 2019 14th ACM/IEEE international conference on human-robot interaction (HRI), 618–619 https://doi.org/10.1109/HRI.2019.8673154

Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0

van Baar, J. M., & FeldmanHall, O. (2021). The polarized mind in context: Interdisciplinary approaches to the psychology of political polarization. *The American Psychologist*, *77*(3), 394–408. https://doi.org/10.1037/amp0000814

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). *Taxonomy of risks posed by language models*. 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229 https://doi.org/10.1145/3531146.3533088

Zhang, L., Pentina, I., & Fan, Y. (2021). Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services. *The Journal of Services Marketing*, *35*(5), 634–646. https://doi.org/10.1108/JSM-05-2020-0162

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2020). The partisan mind: Is extreme political partisanship related to cognitive inflexibility? *Journal of Experimental Psychology: General*, *149*(3), 407–418. https://doi.org/10.1037/xge0000661

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, Y., & Moore, A. (2025). Intuitive judgements towards artificial intelligence verdicts of moral transgressions. *British Journal of Social Psychology*, *64*, e12908. https://doi.org/10.1111/bjso.12908