A Bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgments

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

Yuxin Liu[1,2] (yliu3310@exseed.ed.ac.uk), and Adam Moore[1] (amoore23@exseed.ed.ac.uk)
[1]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK
[2]Centre for Technomoral Futures, Edinburgh Futures Institute, University of Edinburgh, UK

Abstract

Despite substantial progress in artificial intelligence (AI), little is known about people's moral intuitions towards AI systems. Given that politico-moral intuitions often influence judgements in non-rational ways, we investigated participants' willingness to act on verdicts provided by an expert AI system, trust in AI, and perceived fairness of AI as a function of the AI system's (dis)agreement with their pre-existing politico-moral beliefs across various morally contentious issues. Results show belief alignment triggered a willingness to act on AI verdicts but did not increase trust or fairness perception of the AI. This result was unaffected by general AI attitudes. Our findings suggest a disassociation between acceptance of AI recommendations and judgements of trust/fairness of the AI, and that such acceptance is partly driven by alignment with pre-existing intuitions.

Keywords: artificial intelligence; human-AI interaction; moral intuitions; belief alignment; political partisanship

Introduction

In the United States, statistical algorithms have been used to gerrymander district boundaries to reinforce minority control over governments, even when large majorities vote otherwise (Daley, 2016). Although other programmes could detect the use of such manipulation tools and their purposes (Cho & Cain, 2020), would voters or courts trust, accept, and act on such verdicts when their own party stands to lose?

Research and development in artificial intelligence (AI) has attracted significant global attention from industry, academics, and governments (Zhang et al., 2021). In particular, narrow or task-specific AI driven by machine learning algorithms are capable of increasingly sophisticated tasks (e.g., Fagnant & Kockelman, 2015; Gorwa et al., 2020; Wall et al., 2012), which inevitably raises ethical implications (Wallach & Allen, 2009), e.g. amplifying racial and gender biases (Cirillo et al., 2020; Gebru, 2020; Mehrabi et al., 2021; Scheuerman et al., 2020), or misusing algorithms for political gain (Daley, 2016). Given the prevalence of such applications, it is problematic that we lack a coherent account of humans' moral intuitions towards these AI systems and what factors might shape people's willingness to accept or reject assistance from them.

Perception of Artificial Intelligence

While some research into human-AI/machine/algorithm relationships shows an algorithm appreciation effect (Logg et al., 2019; Robinette et al., 2016), people often prefer, trust, and rely more on advice given by human agents than they do robots, machines, or computer-based systems (Dietvorst et al., 2015; Jauernig et al., 2022; Longoni et al., 2019; Önkal et al., 2009; Prahl & van Swol, 2021; Promberger & Baron, 2006; Shaffer et al., 2013). In particular, perceived task characteristics play an important role – trust and comfort with AI increase for automatable or mechanical tasks compared to tasks that require human decisions (Bigman & Gray, 2018; Castelo et al., 2019; Lee, 2018; Schepman & Rodway 2020). Additionally, people are not yet ready to approve AI as capable and accountable moral agents, as shown by the inconsistent evidence on people's attributions of moral norms, permissibility, blame, and accountability to AI versus humans (Banks, 2020; Bonnefon et al., 2016; Hong, 2020; Kahn et al., 2012; Malle et al., 2015; 2019; Shank et al., 2019, 2021; Shank & DeSanti, 2018; Shariff et al., 2017). However, people's acceptance of and trust in AI may be improved by increasing the perceived objectivity of the task performance (Castelo et al., 2019), and limiting AI to an advisory role or emphasising its expertise (Bigman & Gray, 2018), suggesting a potential in future human-AI partnership.

Moral Intuitions and Political Ideologies

The current literature on social perception of AI raises an interesting question: do people hold strong moral intuitions about AI generally, or do their moral judgements about the acceptability of AI vary systematically with their underlying intuitions regarding the domain where the AI is deployed? That is, will people see AI suggestions as a kind of neutral external viewpoint that could potentially cut through divisive issues, or will their intuitions/beliefs about a given topic drive their acceptance/rejection of AI advice?

Whilst there remains debate regarding whether political ideologies or moral intuitions are psychologically more primary (Smith et al., 2017), political ideology can serve as a valuable proxy for predicting a wide range of moral intuitions on various politically charged issues (Hatemi & McDermott, 2016; Hatemi et al., 2019). Recent work in social and political psychology on identity politics and in-out group partisanship shows a kind of information selection that creates highly polarised, self-perpetuating belief systems that interpret identical incoming information to update beliefs in distinctly different ways (Cook & Lewandowsky, 2016; Gaines et al., 2007; Geschke et al., 2019; Jern et al., 2014; Lauderdale, 2016; van Baar & FeldmanHall, 2021). Indeed, people tend to accept or reject incoming information as a function of compatibility between new information and existing ideology/worldview, regardless of, or even at the expense of

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

its factual nature (Brewer, 2012; Flynn et al., 2017; Glinitzer et al., 2021; Hameleers & van der Meer, 2020; Taber & Lodge, 2006). Importantly, this can be better explained by motivated reasoning accounts (Jost et al., 2003, 2017; Jost & Amodio, 2012; Jost & Krochik, 2014; Kahan, 2016a, 2016b; Krochik & Jost, 2011; Moore et al., 2021) than by accounts of effortful rejection of misinformation (Pennycook & Rand, 2019; Roozenbeek & van der Linden, 2019).

These polarising political belief systems are deeply linked to the moral domain, where moral judgements are often the product of, or at least strongly influenced by, seemingly non-rational intuitions, heuristics, or naïve theories, and post hoc effortful reasoning serves an argumentative function to justify one's own views (Baron, 1992, 1995; Haidt, 2001, 2012; Mercier, 2016; Mercier & Landemore, 2012; Mercier & Sperber, 2011; Sunstein, 2005). For example, the five-factor categorisation of moral intuitions, Moral Foundations Theory (Graham et al., 2009, 2011; Haidt & Graham, 2007; Haidt & Hersh, 2001; Haidt & Joseph, 2004; see also Haidt, 2012; Iyer et al., 2012) has often been applied in political contexts: liberals consistently show greater endorsement for care and fairness than conservatives who endorse both individual-centred (care and fairness) and group-binding (loyalty, authority, and purity) foundations more evenly. Thus, we may explore how people react to the deployment of AI in contexts where they have strong, pre-existing moral intuitions based on their political orientation.

The Current Research

By selecting politically polarised topics, we can reliably elicit moral intuitions independent of AI use. In this context, we investigate whether verdicts of potential bias detected by a task-specific AI/algorithm are sufficient evidence to trigger willingness to pre-commit to an investigation. The key manipulation is the intuition or belief (in)compatibility of the AI verdict—does the AI-detected misconduct conform to people's pre-existing intuitions/beliefs? Viewing the AI as a neutral arbiter should increase acceptance of the verdict even when it contradicts pre-existing intuitions. Alternatively, if AI input is subject to context-based motivated reasoning, then its verdicts will be more acceptable when aligned with pre-existing beliefs, and less acceptable when they conflict. Furthermore, trust and fairness perception are common moral judgements about various forms of authorities/experts (de Cremer & Tyler, 2007; Promberger & Baron, 2006), and are both linked to the acceptance of, and reaction to, outcomes (Bianchi et al., 2015; Skitka & Mullen, 2002; Tyler & Degoey, 1996; Tyler & Smith, 1999). Hence, we also examine trust in the AI and perceived fairness of the AI, which have been widely investigated in the field of human-machine interaction (e.g., Castelo et al., 2019; Lee, 2018).

A related question remains: do people have strong, inherent general moral intuitions about AI independent of the context of its usage? Evidence reviewed above indicates largely inconsistent and contradictory judgements about AI use in the society: while some people are inclined to taking advantage of the immense computational power of AI, more are averse to delegating moral decisions that require human judgments to machines. Hence, we include general attitudes towards AI as a covariate to address this point.

The logic is that people may or may not have general moral intuitions about AI itself. If they do, then such intuitions should predict their judgements about AI across contexts. Otherwise, people may instead spontaneously construct moral intuitions about AI as a function of intuition/belief compatibility within a given context. Thus, we predict: (1) increased willingness to accept default actions recommended by AI systems if they align with participants' pre-existing moral/political intuitions, vs. when they do not align; (2) increased trust in the AI when their recommendations align, vs. when they do not align; (3) increased perception of fairness of the AI when their recommendations align, vs. when they do not align; (4) an interaction between the belief alignment effects and political position, with conservatives showing stronger effects than liberals; and (5) the belief alignment effects will remain after the inclusion of both positive and negative general attitudes towards AI. We conducted two experiments (OSF: osf.io/7qjt3): E1 (within-subjects) and E2 (between-subjects), i.e., two samples of subjects received either multiple scenarios across topical foci in E1, or only one scenario in E2. Finding consistent effects across E1 and E2 should increase confidence in the results.

Methods

Participants

Two hundred and two (67 males and 132 females; $M_{\text{age}} = 36.7$ years, $SD_{\text{age}} = 13.36$ years) and 302 native English-speaking adult participants (109 males and 191 females; $M_{\text{age}} = 37.66$ years, $SD_{\text{age}} = 14.09$ years) took part in E1 and E2, respectively (see Procedures below). Testing was conducted online via Qualtrics integrated into the crowdsourcing platform Prolific Academic 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.

Study Design and Materials

We collected data on 1) basic demographics, 2) general attitudes towards AI, and 3) intuitive responses to hypothetical scenarios of judgements made by AI systems, which were presented in a random order. The same study design and materials were used for both E1 & E2, except for different numbers of scenarios participants received.

Demographic Information We collected participants' age, gender, and aspects of political orientations. 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 (1 = *very left-wing/liberal* to 7 = *very right-wing/conservative*).

General Attitudes Towards AI The General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2020) consists of twelve positive attitude items capturing the potential societal and personal benefits of AI utilities (e.g., "I am interested in using artificially intelligent systems in my daily life"; 1 = *strongly disagree* and 5 = *strongly agree*), and eight negative attitude 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*). The negative items were reverse-coded at data collection 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 of the twenty items as one construct.

Hypothetical Scenarios We created hypothetical scenarios with organisations employing reliable expert AI systems to assess statistical anomalies in their everyday operations, and the systems detect a potentially biased human agent (e.g., Table 1). Eight items represent a fully factorial 2 (Context: Left-wing/Liberal or Right-wing/Conservative moral intuitive direction) x 2 (Approve or Reject action taken by the human agent) x 2 (Financial or Judicial domain of the scenario) design. Context indicates an AI verdict presumably 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 content of the scenarios (financial or judicial), 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.

Table 1: Two examples of hypothetical scenarios.

Left-wing/liberal 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*."
Right-wing/conservative context: "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 italicised for the participants.

For each scenario, participants responded to three separate probe questions measuring different aspects of intuitions towards AI 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., investigative actions) based solely on the AI's detection of possible prejudice ("Based on the AI's recommendation, I think that this person in the scenario should be investigated"). Trust in the AI refers to the extent to which participants perceive the AI judgement to be trustworthy ("I trust the AI's judgement in this case"). Perceived Fairness refers to the extent to which they perceived the AI as fair and appropriate ("I believe that the AI is being fair in this case").

Procedures

Eligible participants completed demographics, GAAIS, and scenario(s) in random order, each section on separate pages.

Procedures differed in E1 and E2 only for scenarios. In E1, participants read two pseudo-randomly selected scenarios, such that they were from opposite factorial cells in each topical focus (e.g., Table 1). In E2, participants were shown one random scenario with relevant minimal alterations to the instruction. After each scenario, participants responded to three probes 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. Participants were directed back to Prolific upon successful completion of the study.

Statistical Analysis Plan

All R code and results can be found on OSF. We opted for Bayesian analysis 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.

Fixed effects of context, participant political orientation, and the interaction of the two were entered into the models as main predictors of interest. Means of positive and negative subscales of GAAIS were entered as covariates of interest to account for participants' pre-existing views of AI unrelated to our scenario design. Age was also included as a nuisance covariate to represent basic familiarity with AI. Unique idiosyncrasies within each item, topic, and individual subject were modelled with random intercepts. All the above parameters were used to simultaneously predict Willingness to Act, Trust in AI, and Perceived Fairness of AI, thus controlling for correlations between these variables and generating unique predictive effects for each outcome.

We standardised political views, general AI attitudes, and scenario responses. We then averaged the three aspects of political views to obtain the final measure of participant political position, where higher scores indicate increasing right-wing conservatism. Using the *brms* package (v. 2.15.0; Bürkner, 2017, 2018) in RStudio (v. 4.0.4; R Core Team, 2021), we estimated Bayesian multilevel models to predict all three DVs, with the main pre-registered model containing the predictors of interest, covariates and the nuisance variable specified above. Several reduced versions of the full model were explored. We computed the expected log pointwise predictive density using Bayesian leave-one-out cross validation method ($ELPD_{LOO}$; Vehtari et al., 2017) and the leave-one-out information criterion (LOO-IC). Furthermore, we used Bayes factors (BFs) to quantify the weight of evidence for one model compared to another (Jeffreys, 1948; Kass & Raftery, 1995; Stefan et al., 2019), adopting a slightly more conservative BF interpretation (Kass & Raftery, 1995), where a 2logBF > 10 would suggest "very strong" evidence for a given model against a comparison. The presented results are from the pre-registered (and statistically superior) model. Further results for exploratory models are available on OSF.

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 $\hat{R}s = 1.0$; 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 (Kruschke, 2014; McElreath, 2015), where $\theta \in$ 95% *HDI* would indicate a 95% credibility that the true parameter value lies within this range.

Results

Descriptive Statistics and Correlations

Table 2 displays descriptive statistics, showing consistent distributions across E1 and E2. All average political views were slightly left-leaning, with ratings on social issues being the most liberal, compared to ratings on economic or foreign policy issues. Participants generally held positive attitudes towards utilities and benefits of AI ($\alpha_{E1PosAtt} = \alpha_{E2PosAtt} = 0.88$), while positivity towards the negative affective items were slightly weaker ($\alpha_{E1NegAtt} = 0.83$, $\alpha_{E2NegAtt} = 0.84$), replicating Schepman and Rodway's (2020) results. Scenario responses revealed that participants showed a 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.

Table 3 shows Bayesian Pearson's zero-order correlations. Positive and (reverse-coded) negative attitudes towards AI were correlated, as higher scores on both subscales indicated more positive attitude towards AI. In addition, positive attitudes weakly correlated only with Trust in the AI (r = 0.20, [0.13, 0.28]) and Fairness Perception of the AI (r = 0.21, [0.13, 0.28]) in E1 (Table 3, lower triangle), and only with Willingness to Act (r = 0.11, [0.02, 0.20]) in E2 (Table 3, upper triangle). Negative attitudes were unrelated to any outcome variables. Notably, Trust and Perceived Fairness were more strongly correlated to each other than either was to Willingness to Act in both experiments.

Planned and Exploratory Analyses

Our pre-registered model simultaneously predicted ratings on all three outcome variables (Table 4). In E1, but not E2, positive general AI attitudes predicted more Trust in AI (β = 0.16, [0.04, 0.28], *SE* = 0.06) and greater Perceived Fairness of AI (β = 0.21, [0.09, 0.34], *SE* = 0.06), suggesting those with more positive attitudes towards the utility of AI were more likely to trust and judge the AI as being fair. Ratings on Willingness to Act were negatively predicted by increasing participant political conservatism in E1 (β = -0.15, [-0.29, -0.01], *SE* = 0.07), and by the conservative moral intuitive context in both E1 (β = -0.58, [-0.93, -0.20], *SE* = 0.18) and E2 (β = -0.45, [-0.72, -0.17], *SE* = 0.14), suggesting that conservatism of both participants and the context were related to less willingness to act on AI verdicts of potential

Table 2: Descriptive summaries of measured variables in Experiments 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)	3.00	6.00	3.47 (1.34)	4.00	6.00
Social Issues	3.15 (1.38)	3.00	6.00	3.16 (1.32)	3.00	6.00
Foreign Policy Issues	3.37 (1.34)	4.00	6.00	3.39 (1.40)	4.00	6.00
Mean Political Position	3.30 (1.25)	3.33	5.67	3.34 (1.25)	3.33	6.00
General Attitudes Towards AI (1 = Negative Attitudes, 5 = Positive Attitudes)						
Positive Subscale	3.33 (0.60)	3.33	2.75	3.30 (0.60)	3.33	3.50
Negative Subscale	2.97 (0.65)	3.00	3.25	3.04 (0.69)	3.12	3.75
Responses to Scenarios (1 = *Strongly Disagree*, 5 = *Strongly Agree*)						
Willingness To Act	3.93 (0.92)	4.07	4.00	3.89 (0.93)	4.06	4.00
Trust	3.56 (0.86)	3.62	4.00	3.44 (0.92)	3.69	4.00
Perceived Fairness	3.67 (0.93)	3.94	4.00	3.56 (0.94)	3.78	4.00

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

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

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

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

transgression. These two variables interacted, but only for Willingness to Act (Figure 1) in both E1 (β = 0.30, [0.11, 0.49], *SE* = 0.10) and E2 (β = 0.28, [0.04, 0.51], *SE* = 0.12). Willingness to act on AI judgements increased as a function of belief alignment, but in the opposite direction as predicted—left-wing/liberals showed a much stronger effect than right-wing/conservatives (see Discussion). Nonetheless, the similarity between results of E1 and E2 provides robust support for the predictive power of context and belief alignment on willingness to act.

We compared various reduced versions of the full model (see OSF). In the best models for both experiments, context and its interaction with political position remained predictive of ratings on Willingness to Act regardless of other effects.

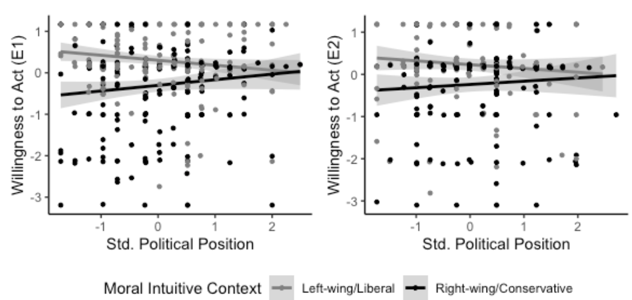

Figure 1: Belief alignment effect for Willingness to Act based on AI verdicts in E1 & E2. Higher standardised scores on political position correspond to increasing conservatism.

Discussion

Both experiments converged on three findings. First, people were generally less willing to act on verdicts of wrongdoing in contexts that matched conservative moral intuitions vs. liberal ones, which might have been skewed by the sample of left-leaning participants, whose politico-moral beliefs were likely violated by the conservative contexts. Second, the belief-alignment effect on participants' willingness to act on AI verdicts trumped general attitudes towards AI, suggesting that people likely have weak to no moral intuitions about AI itself. Rather, judgements about willingness to act on AI advice were instead predominantly driven by whether the AI's recommendation aligned with pre-existing politico-moral intuitions cued by the scenario context, consistent with motivated social cognition needs (Jost, 2017; Jost et al., 2003, 2017; Jost & Amodio, 2012; Jost & Krochik, 2014; Kahan, 2016a, 2016b; Krochik & Jost, 2011). Third, willingness to act on AI advice was not meaningfully related to judgements of trustworthiness or fairness of the AI system itself. This resembles the disjunction between acceptability of outcome (distributive fairness; Ambrose & Arnaud, 2013) vs. fairness perception of the procedure (procedural justice; Cropanzano & Ambrose, 2001) in social justice research, i.e., even if people accept the process as trustworthy and/or fair, they may react unfavourably towards dis-preferred outcomes.

Implication and Future Directions

Several implications of these results come to light. First, we provide empirical evidence that people do not hold strong general moral intuitions towards AI itself. Rather, intuitions towards AI systems seem to be spontaneously constructed, partly driven by a belief alignment effect depending on the intersection of pre-existing intuitions and decision context. Hence, public perception of the acceptability of AI use is likely highly malleable and may be manipulated by framing effects targeting the underlying intuitions associated with different contexts. This clarifies an important distinction between suggestions for advancing human-AI partnership that focus on perceived objectivity of the task (Castelo et al., 2019), versus on presentations of AI itself (e.g., advisory role, expertise or experience; Bigman & Gray, 2018) or humans' control over algorithms (Dietvorst et al., 2018). Framing the setting may thus dominate other means of attempting to shape

Table 4: Full summaries of Bayesian regression fixed effects coefficients for Experiments 1 and 2.

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

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

general AI perception, which will require further normative discussions regarding the ethical design of AI in the future.

Our results also suggest not everyone is equally likely to accept AI recommendations in the face of ideological clashes. Indeed, more extreme ideological beliefs may be associated with stronger biases (cf. van Linden et al., 2021). Further studies should explore satisfaction of AI-produced outcomes (distinct from fairness; van den Bos et al., 1998), confidence in AI decisions (distinct from trust; Earle & Siegrist, 2006), and prompting a view of AI as helpful in provocative settings.

Nonetheless, limitations in our study call for improvement. First, more politically diverse sampling is needed, as our sample of participants may lack “genuine” conservatives, potentially rendering the observed belief alignment effect unreliable for the right end of the continuum. In addition, issue-specific intuitions are not monolithic on either end of the political spectrum, as most people lack ideological coherence (Kalmoe, 2020) and hold moral/political beliefs on some issues that diverge from their self-identified partisan stances (Smith, 2019). While people do tend to have a general political identity that drives affective intuitions (Baldassarri & Page, 2021; Iyengar et al., 2019), the three-item scale we used for overall political orientation may be inadequate for capturing issue-specific beliefs probed by our scenarios of AI use. Future research should use a more expansive instrument to measure specific beliefs (e.g. Everett, 2013), or directly target moral intuitions (Graham et al., 2011), and should explore a broader range of vignettes using topics that have less consensus across the political spectrum, e.g., positive discrimination or affirmative action, or punitive vs. rehabilitative incarceration (Smith, 2019). Moreover, the complexity of our materials may have contributed to a degree of confusion. General AI attitudes' impact may also increase when the AI's role is more salient/causally central to concrete outcomes. To demonstrate the lack of impact of general AI attitudes more rigorously, future studies could compare relatively simple scenarios with and without AI involvement to demonstrate homogeneity of belief-alignment effects.

Conclusion

We studied people's judgements about the willingness to act on an expert AI's detection of potential wrongdoing, trust in the AI, and perceived fairness of the AI across contentious issues. We found politico-moral belief alignment between people and the contexts impacted willingness to follow the course of AI-suggested action, over and above general attitudes towards AI, which is congruent with motivated reasoning. This effect did not promote trust or fairness perception of the AI, indicating a disassociation with willingness to act on the AI's decisions. Further research may investigate the influencing factors in the construction of moral intuitions towards AI generally, and the contexts in which it is to be employed.

Acknowledgments

We thank Sangeet Khemlani, Anne Templeton, and our anonymous reviewers for their valuable feedback on the paper manuscript. Participants were compensated using the postgraduate research student grant from School of Philosophy, Psychology and Language Sciences, University of Edinburgh.

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). Mahwah, NJ: Lawrence Erlbaum Associates.

Baldassarri, D., & Page, S. E. (2021). The emergence and perils of polarization. *Proceedings of the national academy of sciences*, *118*(50), e2116863118.

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.

Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology*, *63*(2), 320–330.

Baron, J. (1995). A psychological view of moral intuition. *The Harvard Review of Philosophy*, *5*(1), 36–40.

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *ArXiv: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.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Brewer, P. R. (2012). Polarisation in the USA: Climate change, party politics, and public opinion in the Obama era. *European Political Science*, *11*(1), 7–17.

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411.

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

Cho, W. K. T., & Cain, B. E. (2020). Human-centered redistricting automation in the age of AI. *Science*, *369*(6508), 1179–1181.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digital Medicine*, *3*(1), 81.

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

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*. Stanford, CA: 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.

Daley, D. (2016). *Ratf**ked: The true story behind the secret plan to steal America's democracy*. New York, NY: Liveright Publishing Corporation.

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.

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.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence. *Journal of Applied Social Psychology*, *36*(2), 383–416.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181.

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.

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.

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: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press/Taylor & Francis Group.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472.

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.

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.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1).

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon Books.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66.

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.

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.

Hatemi, P. K., Crabtree, C., & Smith, K. B. (2019). Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, *63*(4), 788–806.

Hatemi, P. K., & McDermott, R. (2016). Give me attitudes. *Annual Review of Political Science*, *19*(1), 331–350.

Hong, J. W. (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.

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.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, *7*(8), e42366.

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. *Philosophy & Technology*, *35*(1), 2.

Jeffreys, H. (1948). *Theory of probability* (2nd Ed.). Clarendon Press; Oxford University Press.

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology*, *38*(2), 167–208.

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion*, *36*(1), 55–64.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

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 *Advances in motivation science* (Vol. 1, pp. 181–231). Elsevier.

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.

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.

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.

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 (HRI)*, 33-40.

Kalmoe, N. P. (2020). Uses and abuses of ideology in political psychology. *Political Psychology*, *41*(4), 771–793.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

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). New York, NY: Psychology Press.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Boston, MA: Academic Press.

Lauderdale, B. E. (2016). Partisan disagreements arising from rationalization of common information. *Political Science Research and Methods*, *4*(3), 477–492.

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

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.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

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. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (Vol. 95, pp. 111–133). Springer International Publishing.

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan* (1st ed.). Boca Raton, FL: CRC Press/Taylor & Francis Group

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.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

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.

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.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

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.

Peer, E., Rothschild, D. M., Evernden, Z., Gordon, A., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *SSRN Electronic Journal*.

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.

Prahl, A., & van Swol, L. (2021). Out with the humans, in with the machines?: Investigating the behavioral and psychological effects of replacing human advisors with a machine. *Human-Machine Communication*, *2*, 209–234.

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.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, 101–108.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65.

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

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on human-computer interaction*, *4*(CSCW1), 1–35.

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, *33*(1), 108–118.

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.

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.

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.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694–696.

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.

Smith, K. B., Alford, J. R., Hibbing, J. R., Martin, N. G., & Hatemi, P. K. (2017). Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science*, *61*(2), 424–437.

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

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–573.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

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*.

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.

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). New York, NY: Psychology Press.

van Baar, J. M., & FeldmanHall, O. (2021). The polarized mind in context: Interdisciplinary approaches to the psychology of political polarization. *American Psychologist*. Advance online publication.

van den Bos, K., Wilke, H. A. M., & Lind, E. A. (1998). When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology*, *75*(6), 1449–1458.

van Hiel, A., Onraet, E., & de Pauw, S. (2010). The relationship between social-cultural attitudes and behavioral measures of cognitive style: A meta-analytic integration of studies. *Journal of Personality*, *78*(6), 1765–1800.

van Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The paranoid style in American politics revisited: An ideological asymmetry in conspiratorial thinking. *Political Psychology*, *42*(1), 23–51.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE*, *7*(8), e43855.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2021). *Artificial intelligence index report 2021*. Stanford, CA: Human-Centered AI Institute, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf