Public Perception on Authority, Transparency, and Bias of AI in Courts: A Vignette Experiment with Chat Intervention

Hendrik Hüning *Lydia Mechtenberg* *Huyen Nguyen*
Arna Woemmel *

January 30, 2022

Abstract

How do ordinary citizens perceive the increasing influence of artificial intelligence (AI) in society? What significance do they attribute to an ethical design and application of AI? This study investigates the normative attitudes of 2,864 subjects in the UK toward the application of AI as a decision-support tool in criminal courts, using a vignette experiment. We vary the trade-off between discrimination-free versus informed judgments of AI in courts, its degree of decision authority, and the transparency of its code, to study the normative value individuals attribute to these design features of AI. Our results show that transparency is the only factor with a significant impact on participants' attitudes. After the vignette experiment, we implement a real-time randomized chat among participants to investigate potential change of attitudes due to peer arguments. We find that participants become more skeptical about AI usage in courts after chatting with others. This finding highlights the importance of understanding public discourse in shaping attitudes toward AI usage in the public domain.

Keywords: artificial intelligence, courts, judges, algorithms

*Department of Economics, Hamburg University, Von-Melle-Park 5, 20146 Hamburg, Germany, Email addresses of the authors are: Arna.Woemmel@uni-hamburg.de, Huyen.Nguyen-1@uni-hamburg.de, Lydia.Mechtenberg@uni-hamburg.de and Hendrik.Huening@uni-hamburg.de.
We are grateful to the WISO lab of Hamburg University for the outstanding technical assistance. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822590. Any dissemination of results here presented reflects only the authors' view. The Agency is not responsible for any use that may be made of the information it contains. This study received ethical approval from the Institutional Review Board of the University of Hamburg. The experiment is pre-registered at AEA RCT.

1 Introduction

The growing prevalence of big data and sophisticated algorithms has led to rapid application of artificial intelligence (AI) across various high-stake public decision-making contexts. Most of these applications are used as AI-assisted tools, with the goal of supporting decision-makers in making highly consequential judgments by providing predictions about adverse outcomes. Examples of such applications include the assistance of judges in sentencing, pre-trial, and bail decisions (Kleinberg et al. 2018), informing police strategies concerning localities of criminal activity (Shapiro 2017), and hiring based on the candidate's predicted job performance (Li et al. 2020). Since these systems rely on pre-defined functions and complex statistical models, they have the potential to improve accuracy, consistency and fairness in human-decision making, thereby promoting social welfare gains and cost-efficiencies (Kleinberg et al. 2018).

Nevertheless, these applications have been met with several concerns about their ethical ramifications. First, they show potential for reproducing and amplifying societal inequalities. As these algorithms are trained on historical data of human decisions, they are likely to institutionalize societal stereotypes, thereby displaying discriminatory bias towards minority groups (Pager and Shepherd 2008). To counteract such possibilities, technical interventions to achieve bias-free decision making can inevitably induce severe trade-offs, most notably by lowering prediction accuracy (Cowgill and Tucker 2020; Rodolfa et al. 2021). Second, the growing integration of algorithms in decision-making procedures also raises challenges on the extent of human agency. How much decision authority should be attributed to algorithms? How much discretion should be left with the human decision-maker? If certain components of a decision, e.g. the prediction of future outcomes, transferred to algorithms, human control over a final decision outcome becomes increasingly restricted. This does not only raise practical issues in connection with accountability and liability concepts, but also brings up fundamental questions on the role of humans in human-machine decision contexts (Aust 2018; Wagner 2019). Third, the inherent complexity of AI raises issues regarding the transparency of the underlying functioning and data base of the AI. Transparency is necessary to answer how a specific decision is reached and, thus, fostering algorithmic interpretability, accessibility and explainability (Pasquale 2015). Given the complexity of algorithms coupled with vast amount of data, a full disclosure of the AI to the public may not yield higher levels of transparency as most of the citizens may lack the necessary technical skills (Green 2021). It remains inaccessible to the majority of the

population. Therefore one pro-posed measure for improving operational and technical transparency could be achieved by ensuring that the development and maintenance of the algorithm is performed by an objective institution, such as public or federal institutions, serving as a 'trusted auditor' (Pasquale 2015; Shneiderman 2016).

The issues of discrimination and accuracy, decision-authority, and transparency of AI are reflected in the applications of predictive algorithms in criminal courts. Across many states in the US, machine learning algorithms provide predictions of a prisoner's recidivism behavior based on large quantities of data and complex risk models to inform judges in bail and sentencing decision (Kehl and Kessler 2017)s. Often developed by for-profit companies, such algorithms promise to improve justice by increasing consistency across judges and comparable cases (Kahneman et al. 2021). Nevertheless, such non-transparent algorithms have shown notable bias against minority groups, along with a lack of accountability from judges who justified their controversial decisions by means of AI. (Angwin et al. 2016; Ludwig and Mullainathan 2021). Notwithstanding the growing consideration of incorporating such AI-assisted tools into the legal courts across the Western world, the current literature has been largely silent about how private people perceive the ethical design and application of such AI-assisted tools.[1]

The young literature on laypeople's attitudes towards AI shows mixed results, while strongly depending on the decision-context (Castelo et al. 2019). On the one hand, people tend to appreciate and have trust in AI in contexts with decision problems that are objective in nature and that require quantitative analysis (e.g. Logg et al. 2019; Marcinkowski et al. 2020). On the other hand, when it comes to moral tasks and problems that cannot easily be quantified, people prefer human decision-making to algorithms (Bigman and Gray 2018; Gogoll and Uhl 2018) and also perceive it as to be fairer to machine-based decision-making (Lee 2018; Nagtegaal 2021). In general, People's knowledge on AI tends to be limited (Grzymek and Puntschuh 2019). While the circumstances under which people accept or disapprove algorithms have been largely examined, the features of the AI design, especially its ethical attributes, have not received much attention in the empirical literature on people's perception of AI (for an exception see Kieslich et al. 2021). Despite the growing theoretical literature on AI ethics (e.g. Mittelstadt 2019; Morley et al. 2020), it is still unclear how much significance the public attributes to

[1]Most academic discussions are restricted to discussions among scholars, policy-makers, and technical experts (e.g. Cowls and Floridi 2018).

the ethical application and design of AI for decision-making. There is some first research in this domain suggesting that people are especially responsive to algorithmic fairness (Chen et al. 2016; Harrison et al. 2020; Wang 2018). Regarding the relative importance of ethical principles, first results suggest that not a particular attribute is significant for people's acceptance of AI usage in the public decision-making context, but that these systems should comply with several ethical principles simultaneously to receive acceptance within society (e.g. Kieslich et al. 2021). Research in this domain is still rare, yet important considering the increasing influence AI has on our society as well as the various trade-offs an ethical design of AI systems entails (Harrison et al. 2020). Moreover, to the best of our knowledge, public discourse and chat interactions have not been investigated in this domain so far.

This study investigates the public attitudes towards the application and ethical design of predictive algorithms in criminal courts of a heterogeneous sample drawn from the general UK population. Given the crucial role of deliberation in liberal democracies, we also study how discussions with randomly matched peers affect individual opinions. We conducted an online vignette experiment with subsequent real-time chat discussions in October and November 2021. The study contains a 2x2x2 full factorial between-subjects design. The vignette scenarios vary along two dimensions (high and low level) in (i) the trade-off between respecting anti-discrimination constraints and maximizing prediction accuracy with potentially biased training data (discrimination); (ii) the discretion that the judge has in using the AI prediction and the agency attributed to the AI itself (authority), and (iii) in the level of public oversight of the AI (transparency). We documented how individuals change their normative evaluations through deliberation with random peers, using a chat discussion.

Our main findings show an ex-ante positive view towards the usage of AI in courts. Transparency is the only factor with a significant positive impact on these attitudes. Participants prefer transparent algorithms that are under federal oversight to privately developed ones with limited public oversight. Counteracting this ex-ante optimism toward AI, we find a negative deliberation effect of the randomized chat discussions. In particular, the (random) chat-group composition had an influence on attitudes. Participants were more likely to reverse their opinion, the more chat partners with opposing views they encountered in the chat. The opposite holds for chat partners with aligned views. Given that participants with ex-ante optimistic views on AI were in the majority in our sample, this indicates that ceteris paribus, pessimists were more persuasive.

Our study contributes a novel empirical evidence to the primarily theoretical literature on the ethics of AI, particularly on issues of discrimination and accuracy, decision authority, and transparency of AI. Successful algorithmic governance requires the public compliance with and (perceived) legitimacy of government actions. The public support of AI systems used in society is a key factor to ensure legitimacy and compliance (Jacobsohn 1977; Wang 2018). Despite the importance of understanding the public attitudes towards AI in policy domains, this has been largely under-considered in public debates. With this work, we also respond to calls from scholars across disciplines for inclusiveness of societal preferences in the debate around AI (Buhmann and Fieseler 2021). Last but not least, the novel chat component in our research enables us to explore the various factors impacting preference and belief formation, as well as the characteristics of persuasive arguments.

The remainder of this paper is as follows. While Section 2 summarizes the experimental design, Section 3 displays our data. Section 4 discusses our key results. Section 5 provides robustness checks and Section 6 concludes.

2 Experiment Design

We designed 2x2x2=8 hypothetical scenarios of AI usage in criminal courts that participants in our experiment had to read and evaluate (between subjects). For each scenario, we asked our participants to indicate their preference for implementing the particular AI described in the scenario as a decision-support tool in courts within the UK on a Likert scale. Afterwards, we asked them to write down their beliefs about both positive and negative potential consequences for implementing this AI, in two respective free-form text boxes.

Survey responses are context-sensitive and susceptible to framing effects (Zhen and Yu 2016). The setting of criminal courts is particularly suitable for investigating attitudes towards the ethical integration of AI for three reasons. First, while the scenario is a hypothetical case in the UK, its potential application is sufficiently realistic. Countries such as the US have implemented AI in their judicial system and other countries around the world are currently planning similar implementations. As we run a similar scenario in our vignettes, we hope that the internal and external validity of our findings is present (Aguinis and Bradley 2014) Second, while the setting is relatively realistic, we expect that participants have not personally experienced AI in

criminal courts. Hence, this constraints the information environment and choice attributes of the given vignette description, thereby reducing the risk of personal bias in response behavior. Third, each of the three ethical aspects, i.e. discrimination and accuracy, authority, and transparency, is central to judicial decision-making contexts. Therefore, we expect that variation in these factors is sufficiently salient to affect individuals' attitudes. Yet, the application of AI into judicial context is still at a nascent stage, which allows for realistic variations in its ethical and practical attributes.

The study starts with four questions, using 5-point Likert-scales, to elicit the following information from participants: (1) self-perceived familiarity with the topic of AI; (2) their estimation of how familiar the other participants are with AI; (3) their individual positive and negative attitudes towards AI. Afterwards, we introduced our participants to the application and functioning of AI systems with respect to criminal court judgments. The scenario description focuses on early release decisions, which represents a common scenario to judges in the UK. Specifically, judges have to decide whether to grant offenders early release from prison while reducing the risk of future recidivism. The AI system was introduced as a support tool for judges by predicting re-offending probabilities of prisoners, given its data analysis from past court cases and judges' decisions. To ensure common understanding, the case description uses entirely graphical illustrations and simple text captions We also have a comprehension question to check if participants understood the case description.

Next, we informed our participants about potential judicial bias in judgement decisions. In particular, judges can have unobserved personal prejudices and can be susceptible to random factors, which could lead to just or unjust release decisions. We then asked our participants to assess how biased the average judge is.[2] We explained the issue of biased AI to our participants and let them know about the possibility to train the AI on data from all judges' historic cases and decisions. We also elicit their belief on whether they think an AI using all past data would make more or less biased predictions than the average judge. [3] Finally, participants learned about the potential trade-off between maximizing discrimination-free decisions and maximizing accuracy of the AI, which translates to restricting versus not restricting a potentially biased training-data set. While discrimination-free decisions may require the AI to neglect variables such as gender, religion, and ethnicity, maximizing

[2]i.e.using a 5-point Likert-scale: 1 = *always biased* to 5 = *never biased*

[3]i.e. using a 5-point Likert-scale: 1 = *a lot more biased* to 5 = *a lot less biased*

accuracy would require it to use all available information that may have some predictive power, potentially including these sensitive variables. We asked our participants whether an AI trained on a restricted data set would generate more accurate predictions than an AI trained on the unrestricted data set.[4]

We then introduced participants to their randomly assigned vignette scenario. At the beginning, we reminded participants that we are interested in their personal judgment about a scenario and that there is no correct or wrong answer to this. The participants were first given a brief text explanation of AI in criminal courts, which was constant across all scenarios:

In several industrial countries worldwide, criminal courts increasingly use artificial intelligence (AI) to assist judges in early release decisions. The AI predicts how likely prisoners will re-offend if released from prison. Each prisoner is assigned a low, middle, or high re-offending risk score. The prediction considers personal characteristics of the prisoner (e.g. criminal involvement, family/relationships, social exclusion, and attitudes) and is based on sophisticated machine learning models which are linked to a national data base on crime behavior. In general, the prediction of future behavior, including the prisoner's probability of re-offending, is one of the main factors a judge considers when making early release decisions.

The second part of the vignette included the specific scenario. Treatment variations are depicted in Table 1.

Table 1: Factor descriptions in the vignette scenarios

Factor	Level: High		Level: Low	
No Discrimination	ND_h:	Adjusting data base and/or model to ensure a non-discriminatory prediction.	ND_l:	No adjustment to potentially discriminatory effects.
Authority	A_h:	The judge is obliged to incorporate the risk score of reoffending.	A_l:	The judge can decide whether and to what extent the risk score of reoffending is incorporated.
Transparency	T_h:	The AI tool is developed by a federal institution.	T_l:	The AI is developed by a private company.

Notes: The table shows a summary of the textual description of factors in the different experimental scenarios.

[4]i.e. using a 5-point Likert-scale: 1 = *a lot more accurate* to 5 = *a lot less accurate*

After reading the scenario, our participants revealed their personal preference for implementing the described AI in criminal courts. Specifically, they answered the following question using a 5-point Likert-scale: *"How strongly do you agree with the following statement?: The AI, as explained in the scenario, should be implemented in criminal courts."* Afterwards, they replied in two separate free-form text boxes to the following respective questions: *"Would this AI, as described in the scenario, have positive [negative] effects? If yes, why?"*

Next, each participant was randomly matched with two other participants who had evaluated the same vignette scenario to discuss in a real-time messenger chat the issue of AI in courts. We asked participants to remain anonymous throughout the chat discussion. The chat lasted 7 minutes. We again reminded the participants that their contributions may have an impact on real-world policy decisions on the integration of AI systems in UK criminal courts. After the chat discussion, we re-elicited the individual preferences by asking the participants if they would like to change their former (dis-)agreement with the implementation of the described AI in criminal courts.

Furthermore, we asked them in two separate questions with how many chat partners they agreed (disagreed) in the discussion. In case they stated that they disagreed with at least one person in the chat, they were presented an additional question whether they think that they convinced any of their chat partners of their viewpoint. Finally, they were asked how many of their chat partners had convincing arguments. The study ended with standard demographic questions.

Participants were recruited through Prolific with informed consent. All participants were informed of their right to discontinue participation at any time. Prolific maintains a heterogeneous panel of UK citizens. Table A2 and Table A3 in the Appendix present details about the sample demographics. Participants received a fixed payment at the usual Prolific rate (£ 2.63 per participation) and no extra payments. Participants were required to be 18 years of age or older. The median response time was 17.41 min. 65% of recruited participants completed the whole experiment.

Before implementing the study, we ran a pilot session with 25 students in the laboratory at University of Hamburg.The respondents did not indicate any confusion when they had the opportunity to give feedback at the end of the survey. The experiment was entirely computer-based. The program for the experiment was designed using the software oTree (Chen et al. 2016).

The study received IRB approval of Hamburg University on 20th April 2021. Data and associated material are available at [URL Link]. The study was pre-registered on the American Economics Association's registry for randomized control trials (AEA) at https://www.socialscienceregistry.org/trials/8200.

3 Data

Overall, we collected data from 2864 participants. Participants were randomly assigned to one of our eight treatments (scenario variations). Treatments were run in 9 sessions plus 8 resample sessions (See Table A1 for details.). Table 2 to Table A3 present some descriptive characteristics of our sample. Subjects are between 18 and 81 years old, averaging 34 years. Overall, 67% of subjects are female. Most participants have a disposable household income of up to 2000 pounds (75.9%). The majority (49.4%) of participants has a college or university degree. On average, our participants feel that both they themselves and the other participants understand AI well (3.14 and 3.22, respectively). Overall, 20% of participants (525 out of 2687) changed their attitude after the chat interaction.

We also see that positive attitudes towards AI (3.14) are slightly more pronounced than negative attitudes (2.92). Moreover, regarding participants' opinion of whether judges are biased in their decisions, responses are quite equally distributed, averaging 3.04. Participants have a slight tendency towards AI being less biased than the average judge (average 2.62).

Table 2: Summary statistics

No.	Variable	Mean	St. Dev.	Min	Max
1	Understand AI	3.14	0.98	1	5
2	Understand AI (others)	3.22	0.77	1	5
3	Positive Attitudes AI	3.14	0.83	1	5
4	Negative Attitudes AI	2.92	0.77	1	5
5	Judges biased?	3.04	0.72	1	5
6	AI more/less biased?	2.62	1.00	1	5
7	AI Accuracy removed data	3.29	1.07	1	5
8	AI Use Opinion (before Chat)	3.24	1.05	1	5
9	Changed Attitude after Chat (Yes=1)	0.20	0.40	0	1
10	AI Use Opinion (after Chat)	2.89	1.08	1	5
11	Age	34.23	12.03	18	81
12	Share Male	0.33	0.47	0	1
13	Share Female	0.67	0.47	0	1
14	Share Diverse	0.01	0.09	0	1

Notes: The number of observations is 2864 for variables 1, 2 and 4 to 8, 2863 for variable 3, 2687 for variable 9, 525 for variable 10 and 2852 for variables 11 to 14. 177 subjects did not chat (because no chat-group could be formed in 7 minutes) and therefore did not have the chance to chat (variable 9).

4 Results

4.1 Vignette effects on attitudes

First test results of our vignette-variation analysis are shown in Table 3. We find that participants prefer a transparent AI, i.e. they prefer that the AI to be implemented in criminal courts is developed and maintained by a public rather than private institution, with the former allowing for more oversight over the inner workings of the AI than the latter. We do not find a significant difference with regard to non-discrimination and authority, nor with regard to interactions of our factors.

Table 3: Direct effects of vignette variations

Treatment	Mean (AI use opinion)	p-value
A_h	3.21	
A_l	3.27	0.823
T_h	3.35	
T_l	3.14	0.000***
ND_h	3.24	
ND_l	3.25	0.968
$A_h * T_h$	3.33	
A or/and $T = l$	3.21	0.234
$A_h * ND_h$	3.21	
A or/and $ND = l$	3.26	0.823
$T_h * ND_h$	3.31	
$T = l$ or/and $ND = l$	3.22	0.823

Notes: The table displays means of our five-point likert scale question regarding participants opinion about the use of the AI in courts (as described in the scenario). Our eight individual vignette treatments are aggregated to accountability high versus low (A_h, A_l), transparency high versus low (T_h, T_l) and fairness high versus low (ND_h, ND_l). Corresponding interactions are also shown. The last column shows corresponding p-values of MWU-tests. P-values are corrected for the 11 hypotheses in Table 3 and Table 5 using the Holm-method.

In a next step, we investigate our direct vignette effects in the light of individual attitudes towards AI, understanding of AI, and demographics, performing ordered logit regressions. Results are depicted in Table 4.

The effects of our vignette treatments without any additional variables are depicted in column 1. The results resonate with those described earlier.

Table 4: Opinion about AI described in scenario

	(1)	(2)	(3)	(4)
Authority	−0.106	−0.118	−0.216	−0.208
	(0.070)	(0.125)	(0.133)	(0.132)
Transparency	0.344***	0.427***	0.491***	0.480***
	(0.070)	(0.125)	(0.129)	(0.131)
NoDiscrimination	−0.055	0.063	0.071	0.060
	(0.070)	(0.123)	(0.128)	(0.129)
Authority*Transparency		0.050	0.046	0.051
		(0.140)	(0.148)	(0.148)
Authority*NoDiscrimination		−0.025	0.074	0.054
		(0.140)	(0.148)	(0.147)
Transparency*NoDiscrimination		−0.213	−0.170	−0.137
		(0.140)	(0.147)	(0.148)
Diff Understand AI			−0.082*	−0.069
			(0.042)	(0.044)
Attitudes AI (composite)			0.417***	0.410***
			(0.029)	(0.030)
Biased Judges			0.185***	0.152**
			(0.058)	(0.059)
AI more/less biased			−0.558***	−0.559***
			(0.042)	(0.042)
AI non-discriminatory data			0.355***	0.340***
			(0.036)	(0.036)
Legal			−0.392	−0.414
			(0.249)	(0.256)
Obs.	2,864	2,864	2,845	2,838
Controls	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
AIC	7524.817	7528.308	6897.589	6872.434

Note: The table reports ordered logit regressions. Dependent variable is the five-point likert-scale about the opinion of AI as described in the corresponding scenario. *Authority*, *Transparency* and *NoDiscrimination* are dummy variables equal to one for participants that saw the high factor variation and zero otherwise. The *Diff Understand AI* denotes an individual's perceived understanding of AI compared to the average understanding. *Legal* is a dummy equal to one for participants with a legal profession and zero otherwise. The regressions include the following *Controls*: *Age* is a participant's age in years. The variable *Female* (*Diverse*) is a dummy that is equal to one for female subjects (subjects that are diverse) and zero otherwise. The variable *Income* reflects seven household income categories ranging from "less than 500" to "more than 5000". *Education* reflects participants' highest education in six categories. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Highly transparent AI is preferred and leads to a more favourable opinion about using the AI in criminal courts. In column 2 to 4, we gradually add interactions of the vignette-factor variations, prior understanding of and attitudes toward AI, opinions about the relative biases of judges versus AI, and participants' individual socio-demographics as controls.

Unsurprisingly, we find that the opinion about the scenario is positively associated with general attitudes towards AI. We show this in a composite measure that sums positive and (reversed) negative attitudes. Moreover, the more a participant believes that judges are biased (AI is biased), the more (less) she favours the implementation of the AI presented in the scenario. In addition, the more a participant thinks that an AI trained on a reduced data set (dropping data that are most related to prejudices) is accurate, the more likely the will be in favor of implementing the AI described in the scenario.

4.2 Chat effects on attitude change

Effect of the exogenous chat-group composition: Besides private and fully individual opinion formation, we also investigate public opinion formation through deliberation in random chat groups. We start by displaying average opinion changes in reaction to the chat experience. Results are displayed in Table 5.

Table 5: Direct effects of chat

Treatment General chat effect	Mean (AI use opinion)	p-value
Before Chat	3.42	
After Chat	2.89	0.000***
Chat-composition (Chat-partners)	Mean (Opinion change size)	p-value
Both in Favor	-0.01	0.968
One in Favor, on Unsure	-0.04	0.257
Balanced	-0.14	0.823
One Against, one Unsure	-0.23	0.968
Both Against	-0.32	

Notes: The upper part of the table displays means of our five-point likert-scale question regarding participants' opinion about the use of the AI in courts (as described in the scenario). The lower part displays means of opinion changes. The last column shows corresponding p-values of MWU-signed tests for before and after chat (paired sample test) and MWU-ranksum test for the other cases. P-values are corrected for the 11 hypotheses in Table 3 and Table 5 using the Holm-method.

First, we find that participants that had the chance to chat, significantly update their opinion about the use of AI in courts: While the average opinion is slightly more in favor of using AI in criminal courts before the chat (3.42), this becomes slightly more unfavorable after the chat (2.89; Wilcoxon-signed-rank test, p-value: 0.000). Thus, on average participants get more pessimistic about the AI after the chat, regardless of the vignette scenario they have seen. This is the first indication that the chat discussion influences individual attitudes towards the use of AI in courts.

Additionally, chat-group composition has an influence on opinion change. As Table 5 reveals, the more of a participant's two chat partners are initially against the implementation of the AI scenario they have seen, the more negative the update of the participant's opinion about the AI. Average opinion change is effectively zero (-0.01) if both chat partners are initially in favor of implementing the AI. The effect of chat-group composition on opinion change is almost linear by approx. 0.1 units per change in the chat-group composition. Differences between "neighbouring" chat-group compositions are, however, not statistically significant.

Next, we measure the size of the opinion change by calculating the difference between participants' opinions about implementing the AI described in their vignette before and after the chat. We then regress opinion change on the (randomized) chat-group composition of prior opinions on implementing AI in courts (as described in the vignettes), adding further controls. Table 6 displays the results. We find that chat-group composition has a significant influence on participants' opinion update. A participant is the more likely to negatively update her opinion, the more co-players she meets that are in against implementing the AI as described in the scenario (*Co-players Against*). The opposite is true for *Co-players in Favor*.

Interestingly, participants that saw the high factor variation of transparency, i.e. the scenario in which AI is developed by a public institution, are less likely to change their opinion. Thus, also the vignette variation itself has an influence on opinion change. We do not however, find any effects for the other factors or for their interactions.

4.3 Effect of the chat interaction

We now move on to investigating if and how the chat interaction itself, i.e. the text participants exchanged, mediates the effect of our treatment, i.e.

Table 6: Opinion change (after chat)

	(1)	(2)	(3)	(4)	(5)	(6)
Authority	0.233	0.222	0.231	0.232	0.220	0.220
	(0.186)	(0.186)	(0.186)	(0.186)	(0.186)	(0.186)
Transparency	−0.434**	−0.434**	−0.418**	−0.435**	−0.434**	−0.434**
	(0.184)	(0.185)	(0.185)	(0.185)	(0.185)	(0.185)
NoDiscrimination	−0.194	−0.190	−0.183	−0.209	−0.206	−0.206
	(0.193)	(0.194)	(0.194)	(0.195)	(0.196)	(0.196)
Auhtority*Transparency	0.044	0.056	0.032	0.037	0.048	0.048
	(0.216)	(0.217)	(0.217)	(0.217)	(0.218)	(0.218)
Authority*NoDiscrimination	−0.160	−0.138	−0.132	−0.146	−0.123	−0.123
	(0.216)	(0.216)	(0.215)	(0.219)	(0.219)	(0.219)
Transparency*NoDiscrimination	0.338	0.332	0.313	0.342	0.336	0.336
	(0.217)	(0.217)	(0.217)	(0.217)	(0.217)	(0.217)
Co-players in favor minus against	0.316***	0.314***	0.320***			
	(0.046)	(0.046)	(0.046)			
Co-players in Favor				0.250***	0.245***	0.245***
				(0.093)	(0.093)	(0.093)
Co-players Against				−0.392***	−0.393***	−0.393***
				(0.107)	(0.107)	(0.107)
Diff Understand AI		−0.027	−0.028		−0.028	−0.028
		(0.050)	(0.052)		(0.050)	(0.050)
Attitudes AI (composite)		0.019	0.026		0.020	0.020
		(0.034)	(0.035)		(0.034)	(0.034)
Legal		0.006	−0.016		0.005	0.005
		(0.323)	(0.318)		(0.323)	(0.323)
Obs.	2,654	2,647	2,641	2,654	2,647	2,641
Controls	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	4072.102	4060.719	4056.196	4073.372	4061.929	4057.394

Note: The table reports ordered logit regressions. Dependent variable is the difference between the opinion about AI before and after chat. *Authority*, *Transparency* and *NoDiscrimination* are dummy variables equal to one for participants that saw the high factor variation and zero otherwise. The variable *Co-players in Favor minus Against* denotes the chat composition an individual faces (in favor minus against AI-opinions of chat-partners). Furthermore, *Co-players In Favor* (*Co-players Against*) denotes the number of chat partners ex-ante in favor (against) AI. *Diff Understand AI* denotes an individual's perceived understanding of AI compared to the average understanding. *Legal* is a dummy equal to one for participants with a legal profession and zero otherwise. Control variables are the same as described in Table 6. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

the chat composition of co-players' priors, on opinion change.

There are several potential pathways in how the co-players' prior attitudes towards the scenario affect a participant's opinion change through the chat discussion. First, co-players might spread bad or good sentiment during the chat discussion that in turn affects an individual's opinion change. Second, co-players might express a strong policy position regarding the use of AI in courts that in turn affects an individual's opinion. We test both potential pathways with a mediation analysis.

We measure an individual's sentiment with a dictionary approach, i.e. we sum all words that are associated with positive and negative sentiment and normalize this frequency by the total number of words an individual wrote during the chat.[5] Moreover, we measure an individual's policy position by the Wordscores model (Laver et al. 2003). Wordscores is a scaling technique that is based on reference texts, where the policy position is known and "virgin" texts for which the model predicts the policy position based on what it learned from the distribution of words in the reference texts. In our case, we use the positive and negative text-box messages as our reference texts. Thus, wordscores learns word distributions from these two types of text and predicts for each participant's chat-text a score. The score reflects how much the text messages of that participant reflects the word distribution from the positive or negative textboxes and therefore of how strongly a positive or negative position with regard to the scenario was expressed during the chat.

First, we investigate if a participant's co-players chat contributions are influenced by their prior attitudes towards the scenario. Results are depicted in Table 7. We see that the more of participant i's co-players are ex-ante in favor (against) of the scenario, the stronger their positive (negative) sentiment. The same is true for their expressed policy position: The more co-players a priori in favor (against) the the scenario, the stronger their positive (negative) policy position. These findings highlight the influence of our exogenous in-chat treatment variation, i.e. the chat-group composition, on the chat discussion. Since we found that the chat composition has a direct effect on opinion change, sentiment and policy positions, the chat composition might also have an indirect effect on opinion change, mediated by its effect on sentiment or the strength of expressed policy positions.

[5]For the sentiment analysis, the polarity function of the qdap package in R is used (Hu and Liu 2004) that also implements a weighting scheme of positive and negative expressions.

Table 7: Co-players' sentiment and policy positions in chat by priors

	Sentiment			Policy Position		
Co-player	Mean	p-value	p-value (0 to 2)	Mean	p-value	p-value (0 to 2)
In Favor						
0	-0.39			-0.47		
1	0.09	0.00***		-0.02	0.00***	
2	0.46	0.00***	0.00***	0.42	0.00***	0.00***
Against						
0	0.29			0.24		
1	-0.11	0.00***		-0.24	0.00***	
2	-0.64	0.00***	0.00***	-0.71	0.00***	0.00***

Notes: The left panel of the table displays means of co-players' sentiment (measured by a dictionary approach) by their prior attitude towards the scenario. The last two columns show corresponding p-values of MWU-tests. The right panel shows the same for policy positions.

As a first indication of a potential mediating effect of sentiment and policy positions expressed in the chat, we add those measures (of the co-players) to our regressions from before investigating opinion change. Results are depicted in Table 8. We see that co-players' sentiment does not affect a participant's opinion change. Co-players policy positions, however, exhibit a sizeable effect on opinion change beyond the effect of co-players' prior attitudes. Moreover, we find that the size of the effect of co-players' prior attitudes (0.320) is significantly reduced after adding co-players' expressed policy position (Sobel's test, p-value 0.034**). This is a first indication that part of the effect of the randomized chat-group composition is mediated by how strongly the co-players express their policy positions in the chat.

As further evidence of this mediation effect, we perform a causal mediation analysis. For an illustration of the mediation effect and its underlying assumptions, see Appendix B. In the following, we empirically investigate if such a mediation effect exists, using the methods proposed in Imai et al. (2010a) and Imai et al. (2010b).[6] As a first step, we formulate the outcome and mediator model as

$$\Delta Opinion_i = \alpha + \beta Priors_i^{CP} + \delta Policypos_i^{CP} + \gamma X_i + \epsilon_i \quad (1)$$

$$Policypos_i^{CP} = \lambda + \theta Priors_i^{CP} + \phi X_i + \eta_i, \quad (2)$$

[6]We use the *mediation* package in R (Tingley et al. 2014) that implements these methods.

Table 8: Opinion change (potential mediators added)

	(1)	(2)	(3)	(4)	(5)	(6)
Authority	0.231	0.236	0.237	0.230	0.235	0.236
	(0.186)	(0.186)	(0.185)	(0.186)	(0.186)	(0.185)
Transparency	−0.418**	−0.420**	−0.411**	−0.418**	−0.420**	−0.411**
	(0.185)	(0.185)	(0.184)	(0.185)	(0.186)	(0.184)
NoDiscrimination	−0.183	−0.191	−0.182	−0.199	−0.206	−0.197
	(0.194)	(0.194)	(0.193)	(0.196)	(0.195)	(0.194)
Authority*Transparency	0.032	0.041	0.063	0.024	0.033	0.055
	(0.217)	(0.216)	(0.216)	(0.218)	(0.217)	(0.218)
Authority*NoDiscrimination	−0.132	−0.149	−0.164	−0.118	−0.135	−0.150
	(0.215)	(0.214)	(0.214)	(0.218)	(0.217)	(0.217)
Transparency*NoDiscrimination	0.313	0.331	0.318	0.317	0.335	0.322
	(0.217)	(0.219)	(0.217)	(0.217)	(0.219)	(0.217)
Co-players in Favor minus Against	0.320***	0.297***	0.288***			
	(0.046)	(0.047)	(0.047)			
Co-players Sentiment		0.071			0.071	
		(0.048)			(0.048)	
Co-players Policy position			0.100***			0.100***
			(0.035)			(0.035)
Sentiment		0.001	0.010		0.002	0.010
		(0.060)	(0.060)		(0.060)	(0.060)
Policy position		0.088**	0.078**		0.087**	0.077**
		(0.039)	(0.039)		(0.039)	(0.039)
Co-players in Favor				0.250***	0.229**	0.222**
				(0.094)	(0.094)	(0.094)
Co-players Against				−0.400***	−0.375***	−0.365***
				(0.107)	(0.108)	(0.108)
Diff Understand AI	−0.028	−0.021	−0.025	−0.029	−0.022	−0.026
	(0.052)	(0.052)	(0.052)	(0.051)	(0.052)	(0.052)
Attitudes AI (composite)	0.026	0.016	0.017	0.026	0.017	0.018
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Legal	−0.016	−0.010	0.001	−0.018	−0.011	−0.0001
	(0.318)	(0.319)	(0.320)	(0.318)	(0.319)	(0.319)
Obs.	2,641	2,641	2,641	2,641	2,641	2,641
Controls	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	4056.196	4055.847	4051.612	4057.394	4057.080	4052.878

Note: The table reports ordered logit regressions. Dependent variable is the difference between the opinion about AI before and after chat. Independent variables are the same as described in Table 6. The variable *(Co-players) Policy position* measures how strongly (co-players) policy positions towards the scenario is expressed in the chat. *(Co-players) Sentiment* measures how strongly (co-players) positive or negative sentiment is expressed in the chat discussion. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

where X contains all covariates that are used as control variables in our opinion change regressions from Table 8. The superscript CP indicates variables of participant i's coplayers. In order to estimate the indirect and direct effect, we simplify the chat-group composition to a binary variable. Thus, $Priors_i^{CP}$ is equal to one if participant i encounters at least one co-player with positive prior attitudes and zero otherwise. These regression models are subsequently used to estimate if there is an indirect causal effect from co-players' prior attitudes on i's opinion change through the strength of policy positions they express during the chat.

Results of a mediation analysis with linear outcome and mediation equations is presented in Table 9. We find that both the average causal mediation effect (ACME) and the average direct effect (ADE) are highly significant. While the size of the direct effect is dominating, there is a substantial role of the mediation effect. In other words, while the prior attitudes of co-players affect a participant's opinion change, the strength of the policy positions they express in the chat additionally changes their opinion. The sign is positive, indicating the stronger the co-players' positive policy position, the more likely they convince the participants in the chat to change her attitude in a positive direction, i.e. being more in favor of the scenario than before. For a sensitivity analysis of this effect see Appendix B.

Table 9: Causal Mediation Analysis

Effect	Estimate	CI lower	CI upper	p-value
ACME	0.011	0.003	0.020	0.010***
ADE	0.149	0.099	0.200	0.000***
Total Effect	0.161	0.110	0.220	0.000***
Prop. Mediated	0.071	0.016	0.140	0.000***

Notes: Confidence intervals are obtained with nonparametric bootstrap using the percentile method. Sample size used: 2641. Simulations: 1000. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

5 Robustness Checks

We perform the following robustness checks in order to show the validity of our results. First, for the regressions presented in Table 4, we used our 5-point Likert scale that measures the attitude towards the scenario as the dependent variable. As Bloem (2021) and others highlight, the use of ordinal

scales in such regressions can be problematic. Therefore, we show, with the solution proposed in Bloem (2021) that our results are sign robust to a range of plausible monotonic increasing transformations of the ordinal scale. We create a dummy variable that is equal to one for positive attitudes towards the scenario and zero otherwise and use it instead of the 5-point Likert scale. Results, depicted in Table C1, show that our results are robust to this transformation of the dependent variable.

Second, participants needed very different amount of time to complete the survey. Some rushed through the survey and (potentially) paid very little attention to its content. Others needed long (potentially) because they had difficulties understanding its content. On order to show that those participants do not affect our results, we remove those who belong to the 10% fastest or the 10% slowest participants and re-estimate our regressions. Results, depicted in Table C2 to Table C5, show that our results remain highly robust in this respect.

Third, at the beginning of the experiment we introduce participants to how Artificial Intelligence can be used at courts. After this introduction, we check with a comprehension question if they read and understood those introductions. Around 6.4% answered this question incorrectly suggesting that either they did not read the introduction or did not understand it. As Table C6 and Table C7 indicate, our results are robust to the exclusion of those participants.

6 Conclusion

In our study, we find that people show a slightly favorable attitude toward the usage of AI as decision support in criminal court decisions, in particular if the AI is provided by a public rather than private firm and access to the code is freely granted (transparent AI). Other factors such as discrimination and authority do not affect prior attitudes. A sizeable minority changes their opinion after a chat with random partners, becoming more pessimistic after the chat. The prior attitudes of chat partners seem affect opinion change after the chat. This effect is partly mediated by expressed policy positions of chat partners during the chat discussion.

By showcasing public attitudes and deliberation, this study helps bridging the dialogue between technologists and social scientists in the design of algorithms which entails important public consequences. Our findings are

especially relevant for policy-makers in AI system regulations, taking into account not only the academic and theoretical debate discussions but also the diverse views of private citizens. For developers, such insights into how the general public perceives the integration of AI into society and its ethical components could facilitate them with developing more inclusive AI systems, especially given the complex trade-offs in connection with ethical AI.

We acknowledge several limitations of our study. First, in our experimental design we implement three specific and distinct features of AI usage in courts that are of ethical relevance. The vignette scenarios do not capture further relevant features, e.g. distributive fairness, privacy, and liability. Our experimental design serves as a starting point, and we believe it is important to test further ethical concepts and human centered values. Second, our sample is not representative of the general population. While we aimed at a diverse study sample by avoiding potential selection-effects, the sample shows a comparably high education level as well as imbalances in terms of race and professional background. Moreover, one should keep in mind that this study only captures participants from the UK. The findings should be taken more cautiously in an international context. It is not clear whether cross-cultural results would be similar, also given substantial differences in terms of technological acceptance and judicial acceptance across countries.

Hence, this study opens up several directions for future research. In addition to empirically exploring public attitudes towards additional features of AI that are ethically relevant, future experiments could investigate different levels of algorithmic authority including fully automated decision-making in courts or other public domains. Second, future work could investigate how people perceive the application of AI as decision tool for high-impact decision-making in other contexts, such as university admissions or hiring, where the decision at hand is more familiar to the participants. Third, further research could examine why algorithmic transparency, i.e. federal oversight of the algorithm, plays such an important role for public attitudes towards AI. Why do people attribute so much importance to the maintenance and oversight by the state? Why do they not consider the factors of algorithmic bias and the allocation of decision rights to an extent as proposed by the theoretical debates? Fourth, a long-term study to investigate the temporal persistence of attitudes would be interesting. Finally, it is important to consider how to incorporate the general public's views into algorithmic decision-making. In the future, in order to test for generalizability of our findings, we aim to collect data from further representative sample pools and externally valid field settings.

References

Aguinis, Herman and Kyle J. Bradley. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods*, 17(4):351–371, 2014.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 23(2016):139–159, 2016.

Aust, Helmut Philipp. Undermining human agency and democratic infrastructures? the algorithmic challenge to the universal declaration of human rights. *AJIL Unbound*, 112:334–338, 2018.

Bigman, Yochanan E and Kurt Gray. People are averse to machines making moral decisions. *Cognition*, 181:21–34, 2018.

Bloem, Jeffrey R. How much does the cardinal treatment of ordinal variables matter? an empirical investigation. *Political Analysis*, page 1–17, 2021.

Buhmann, Alexander and Christian Fieseler. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, 64:101475, 2021.

Castelo, Noah, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

Chen, Daniel L., Martin Schonger, and Chris Wickens. otree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

Cowgill, Bo and Catherine E. Tucker. Economics, fairness and algorithmic bias. *in preparation for: Journal of Economic Perspectives*, 2020.

Cowls, Josh and Luciano Floridi. Prolegomena to a white paper on an ethical framework for a good ai society. *Available at SSRN 3198732*, 2018.

Gogoll, Jan and Matthias Uhl. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74:97–103, 2018.

Green, Ben. The flaws of policies requiring human oversight of government algorithms. *Available at SSRN*, 2021.

Grzymek, Viktoria and Michael Puntschuh. What europe knows and thinks about algorithms results of a representative survey. bertelsmann stiftung eupinions february 2019. 2019.

Harrison, Galen, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 392–402, New York, NY, USA, 2020. Association for Computing Machinery.

Hu, Minqing and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, page 755–760. AAAI Press, 2004. ISBN 0262511835.

Huber, Martin. Mediation analysis. In *Handbook of Labor, Human Resources and Population Economics*, pages 1–38. Springer, 2020.

Imai, Kosuke, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010a.

Imai, Kosuke, Luke Keele, and Dustin Tingley. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1): 51–71, 2010b.

Jacobsohn, Gary J. Citizen participation in policy-making: the role of the jury. *The Journal of Politics*, 39(1):73–96, 1977.

Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein. *Noise: a flaw in human judgment*. Little, Brown, 2021.

Kehl, Danielle Leah and Samuel Ari Kessler. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. 2017.

Kieslich, Kimon, Birte Keller, and Christopher Starke. Ai-ethics by design. evaluating public perception on the importance of ethical design principles of ai. *arXiv preprint arXiv:2106.00326*, 2021.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

Laver, Michael, Kenneth R Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

Lee, Min Kyung. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

Li, Danielle, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Technical report, National Bureau of Economic Research, 2020.

Logg, Jennifer M, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

Ludwig, Jens and Sendhil Mullainathan. Fragile algorithms and fallible decision-makers: Lessons from the justice system. *Journal of Economic Perspectives*, 35(4):71–96, 2021.

Marcinkowski, Frank, Kimon Kieslich, Christopher Starke, and Marco Lünich. Implications of ai (un-)fairness in higher education admissions: The effects of perceived ai (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 122–130, New York, NY, USA, 2020. Association for Computing Machinery.

Mittelstadt, Brent. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1:501–507, 2019.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26:2141–2168, 2020.

Nagtegaal, Rosanna. The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, 38(1):101536, 2021.

Pager, Devah and Hana Shepherd. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, 34(1):181–209, 2008.

Pasquale, Frank. *The black box society*. Harvard University Press, 2015.

Rodolfa, Kit, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3:896–904, 2021.

Shapiro, Aaron. Reform predictive policing. *Nature news*, 541(7638):458, 2017.

Shneiderman, Ben. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48):13538–13540, 2016.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Kosuke Imai, and Luke Keele. mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38, 2014.

Wagner, Ben. Liable, but not in control? ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, 11(1):104–122, 2019.

Wang, A.J. Procedural justice and risk-assessment algorithms. *Available at SSRN 3170136*, 2018.

Zhen, Shanshan and Rongjun Yu. All framing effects are not created equal: Low convergent validity between two classic measurements of framing. *Scientific Reports*, 6(1):1–9, 2016.

Appendix A

Table A1: Treatments and Sessions

Vignette Treatment	Obs.	Session Dates (Year 2021)
$A_hT_lF_l$	383	20/10, 25/10, 15/11
$A_hT_lF_h$	363	26/10, 15/11
$A_hT_hF_l$	353	27/10, 16/11
$A_hT_hF_h$	354	28/10, 16/11
$A_lT_lF_l$	353	29/10, 17/11
$A_lT_lF_h$	350	02/11, 17/11
$A_lT_hF_l$	358	03/11, 18/11
$A_lT_hF_h$	350	04/11, 23/11

Notes: The table displays the number of participants per treatment and detailed dates on the session timing.

Table A2: Disposable Household Income

Category	Income level	Frequency	Percentage
1	less than 500	755	26.5%
2	500-1000	769	27.0%
3	1000-2000	636	22.4%
4	2000-3000	320	11.2%
5	3000-4000	165	5.8%
6	4000-5000	77	2.7%
7	more than 5000	123	4.3%

Notes: The table displays frequencies and percentages of responses across seven disposable net household income levels. Income is in British pounds. 19 out of 2864 participants did not answer the question.

Table A3: Education distribution

Category	Education level	Frequency	Percentage
1	Primary School	2	0.1%
2	Secondary School (up to 16 years)	199	7.0%
3	Higher Education (A-levels, BTEC, etc.)	670	23.5%
4	College or University	1407	49.4%
5	Post-graduate Degree	560	19.6%
6	Prefer not to say	12	0.4%

Notes: The table displays frequencies and percentages of responses across six education categories. 14 out of 2864 participants did not answer this question.

Table A4: Correlation matrix

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. AI opinion change (dummy)	1.00															
2. Age	-0.10	1														
3. Understand AI	-0.05	-0.08	1													
4. Understand AI (others)	0.00	-0.03	0.45	1												
5. Positive Attitudes	-0.01	-0.1	0.25	0.11	1											
6. Negative Attitudes	0.02	0.08	-0.08	-0.02	-0.61	1										
7. Biased Judges?	0.02	-0.15	0.06	0.02	0.04	0	1									
8. AI more/less biased	0.02	-0.03	0.04	-0.03	-0.03	0.06	0.05	1								
9. AI non-discriminatory data	0.07	-0.1	0.01	0.04	0.06	-0.06	0.04	0.02	1							
10. AI opinion 1	0.07	-0.11	0.03	0.04	0.25	-0.25	0.06	-0.26	0.19	1						
11. AI opinion 2		0.04	-0.04	0.07	0.14	-0.04	-0.06	-0.09	0.05	-0.11	1					
12. Female	0.05	-0.1	-0.3	-0.04	-0.16	0.07	0.1	0	0.06	0.04	0.02	1				
13. Male	-0.06	0.11	0.3	0.04	0.16	-0.07	-0.11	-0.01	-0.07	-0.04	-0.02	-0.98	1			
14. Diverse	0.01	-0.05	0.03	0	0	0	0.06	0.06	0.03	0	0	-0.13	-0.06	1		
15. Infavor (co-player)	-0.06	-0.05	0.01	0	-0.01	0.02	-0.01	-0.02	-0.02	0.02	0.19	-0.05	0.05	0.01	1	
16. Against (co-player)	0.06	0.04	-0.02	-0.01	0.04	-0.04	-0.01	-0.01	0	-0.01	-0.23	0.01	-0.01	-0.01	-0.6	1

Notes: The table presents Spearman's correlations among all variables used in the regressions.

Appendix B

Direct and indirect path

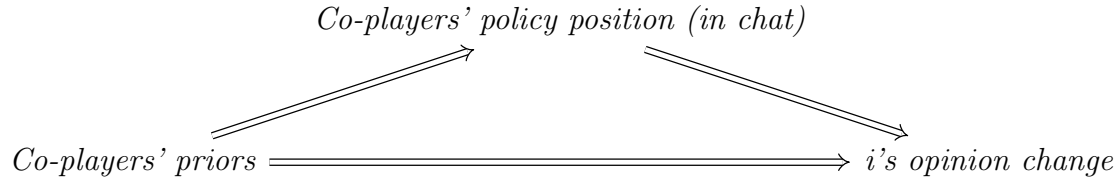

Figure B1: Potential Mediation

We identify a causal mediation effect based on the following four assumptions: There must not be confounders between treatment and outcome relationship (Assumption 1), between mediator and outcome relationship (Assumption 2), or between treatment and mediator relationship (Assumption 3). In addition, there must not be confounders affected by the treatment between mediator and outcome relationship (Assumption 4). These assumptions are also known as sequential ignorability or sequential independence assumptions (Huber 2020).

Assumptions 1 and 3 are met because our treatment is randomized. Assumption 2 is met, if there are no other in-chat factors that confound the relationship between co-players' expressed policy position and participant i's opinion change. As we control for the potential role of participant i's own expressed policy position and sentiment on her co-players communication, confounding effect can be excluded. Similar arguments can be put forward for assumption 4.

As this mediation analysis rests on the assumptions of no other factors confounding this relationship, we perform a sensitivity analysis (also proposed in Tingley et al. 2014). The sensitivity analysis has the following intuition: If confounding factors were missing form the outcome and mediation equation in (1) and (2), their error terms (ϵ_i and η_i) would be correlated. The sensitivity analysis investigates how strong this correlation (denoted as ρ) would have to be so that the mediation effect is zero. In our case, the correlation between outcome and mediator model is 6.401144e-17. The sensitivity analysis reveals that only if confounding factors lead to a correlation of $\rho = 0.05$ or higher, the causal mediation effect would be zero or even reverse its sign (compare Figure B2).

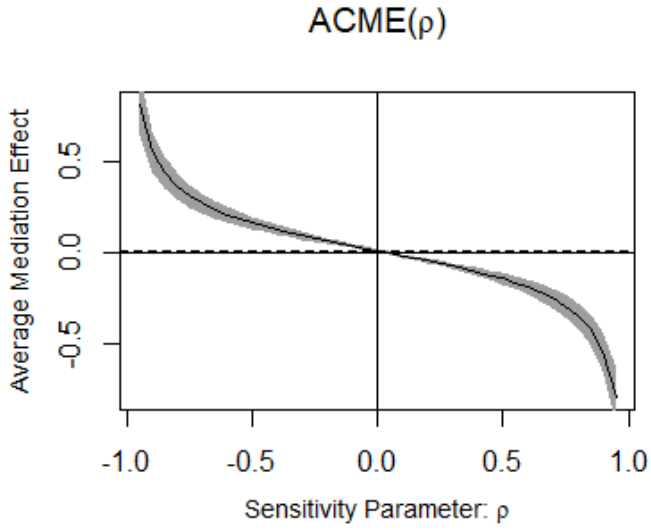

Figure B2: Sensitivity Analysis - ACME by potential values of ρ

Appendix C

Table C1: Opinion about AI described in scenario (Dependent variable as Dummy)

	(1)	(2)	(3)	(4)
Authority	−0.082	−0.090	−0.144	−0.150
	(0.076)	(0.126)	(0.137)	(0.138)
Transparency	0.310***	0.425***	0.482***	0.465***
	(0.076)	(0.129)	(0.138)	(0.140)
NoDiscrimination	−0.109	−0.053	−0.036	−0.051
	(0.076)	(0.130)	(0.138)	(0.140)
Authority*Transparency		−0.048	−0.080	−0.063
		(0.151)	(0.165)	(0.166)
Authority*NoDiscrimination		0.065	0.149	0.136
		(0.151)	(0.165)	(0.166)
Transparency*NoDiscrimination		−0.181	−0.138	−0.104
		(0.151)	(0.165)	(0.166)
Diff Understand AI			−0.072	−0.055
			(0.044)	(0.047)
Attitudes AI (composite)			0.387***	0.380***
			(0.032)	(0.033)
Biased Judges			0.185***	0.128**
			(0.058)	(0.059)
AI more/less biased			−0.496***	−0.506***
			(0.044)	(0.044)
AI non-discriminatory data			0.301***	0.280***
			(0.037)	(0.037)
legal			−0.234	−0.266
			(0.273)	(0.282)
Constant	0.092	0.052	−2.616***	−2.063***
	(0.073)	(0.096)	(0.317)	(0.408)
Obs.	2,864	2,864	2,845	2,838
Controls	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
AIC	3,942.264	3,946.522	3,534.067	3,502.157

Note: The table reports binary logit regressions. Dependent variable is a dummy variable that is equal to one if the attitudes towards the scenario where positive and zero otherwise. Independent variables are the same as described in Table 4. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C2: Opinion about AI described in scenario (Removing the fastest 10% of participants)

	(1)	(2)	(3)	(4)
Authority	−0.100	−0.059	−0.139	−0.128
	(0.074)	(0.131)	(0.138)	(0.138)
Transparency	0.345***	0.414***	0.524***	0.514***
	(0.074)	(0.131)	(0.135)	(0.136)
NoDiscrimination	−0.043	0.067	0.081	0.069
	(0.074)	(0.128)	(0.133)	(0.133)
Authority		0.004	−0.046	−0.046
		(0.147)	(0.155)	(0.155)
Authority*NoDiscrimination		−0.083	0.020	−0.001
		(0.147)	(0.155)	(0.155)
Transparency*NoDiscrimination		−0.139	−0.114	−0.078
		(0.147)	(0.155)	(0.155)
Diff Understand AI			−0.080*	−0.075
			(0.044)	(0.046)
Attitudes AI (composite)			0.429***	0.420***
			(0.031)	(0.032)
Biased Judges			0.138**	0.102*
			(0.059)	(0.060)
AI more/less biased			−0.564***	−0.569***
			(0.045)	(0.045)
AI non-discriminatory data			0.352***	0.340***
			(0.038)	(0.038)
Legal			−0.338	−0.373
			(0.280)	(0.286)
Obs.	2,577	2,577	2,564	2,558
Controls	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
AIC	6811.767	6816.545	6247.250	6222.729

Note: The table reports ordered logit regressions. Dependent variable is the five-point likert-scale about the opinion of AI as described in the corresponding scenario. Independent variables are the same as described in Table 4. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C3: Opinion about AI described in scenario (Removing the slowest 10% of participants)

	(1)	(2)	(3)	(4)
Authority	−0.115	−0.088	−0.223	−0.229
	(0.074)	(0.134)	(0.141)	(0.141)
Transparency	0.313***	0.401***	0.441***	0.428***
	(0.074)	(0.134)	(0.136)	(0.138)
NoDiscrimination	−0.081	0.060	0.058	0.049
	(0.074)	(0.130)	(0.136)	(0.137)
Authority*Transparency		0.028	0.082	0.100
		(0.147)	(0.155)	(0.156)
Authority*NoDiscrimination		−0.081	0.042	0.029
		(0.147)	(0.155)	(0.155)
Transparency*NoDiscrimination		−0.200	−0.159	−0.129
		(0.147)	(0.154)	(0.155)
Diff Understand AI			−0.087**	−0.079*
			(0.043)	(0.045)
Attitudes AI (composite)			0.413***	0.408***
			(0.030)	(0.032)
Biased Judges			0.199***	0.173***
			(0.062)	(0.063)
AI more/less biased			−0.540***	−0.541***
			(0.045)	(0.045)
AI non-discriminatory data			0.331***	0.315***
			(0.038)	(0.039)
Legal			−0.491**	−0.495*
			(0.250)	(0.256)
Obs.	2,577	2,577	2,563	2,557
Controls	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
AIC	6733.952	6737.741	6204.425	6189.609

Note: The table reports ordered logit regressions. Dependent variable is the five-point likert-scale about the opinion of AI as described in the corresponding scenario. Independent variables are the same as described in Table 4. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C4: Opinion change (Removing the fastest 10% of participants)

	(1)	(2)	(3)	(4)	(5)	(6)
Authority	0.198	0.204	0.203	0.196	0.201	0.200
	(0.192)	(0.192)	(0.191)	(0.192)	(0.191)	(0.191)
Transparency	−0.391**	−0.396**	−0.386**	−0.393**	−0.397**	−0.388**
	(0.192)	(0.192)	(0.191)	(0.192)	(0.192)	(0.191)
NoDiscrimination	−0.231	−0.248	−0.237	−0.248	−0.264	−0.253
	(0.199)	(0.199)	(0.198)	(0.201)	(0.200)	(0.200)
Authority*Transparency	−0.039	−0.034	−0.007	−0.045	−0.041	−0.014
	(0.223)	(0.222)	(0.222)	(0.224)	(0.223)	(0.224)
Authority*NoDiscrimination	−0.069	−0.081	−0.097	−0.051	−0.064	−0.080
	(0.220)	(0.219)	(0.219)	(0.224)	(0.223)	(0.223)
Transparency*NoDiscrimination	0.283	0.314	0.295	0.287	0.317	0.298
	(0.223)	(0.224)	(0.222)	(0.223)	(0.224)	(0.222)
Co-players in Favor minus Against	0.327***	0.297***	0.295***			
	(0.048)	(0.048)	(0.048)			
Co-players Sentiment		0.098*			0.098*	
		(0.051)			(0.051)	
Co-players Policy position			0.096***			0.096***
			(0.037)			(0.037)
Sentiment		0.013	0.029		0.014	0.030
		(0.061)	(0.060)		(0.061)	(0.060)
Policy position		0.086**	0.074*		0.085**	0.073*
		(0.043)	(0.043)		(0.043)	(0.043)
Co-plaers in Favor				0.255***	0.226**	0.227**
				(0.096)	(0.096)	(0.096)
Co-players Against				−0.409***	−0.379***	−0.374***
				(0.112)	(0.113)	(0.113)
Diff Understand AI	−0.019	−0.011	−0.015	−0.021	−0.012	−0.016
	(0.053)	(0.054)	(0.053)	(0.053)	(0.054)	(0.053)
Attitudes AI (composite)	0.030	0.020	0.020	0.031	0.021	0.021
	(0.036)	(0.036)	(0.036)	(0.036)	(0.037)	(0.037)
Legal	−0.123	−0.111	−0.102	−0.123	−0.112	−0.104
	(0.306)	(0.308)	(0.308)	(0.306)	(0.308)	(0.308)
Obs.	2,402	2,402	2,402	2,402	2,402	2,402
Controls	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	3713.823	3711.783	3710.506	3715.045	3713.020	3711.793

Note: The table reports ordered logit regressions. Dependent variable is the difference between the opinion about AI before and after chat. Independent variables are the same as described in Table 8. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C5: Opinion change (Removing the slowest 10% of participants)

	(1)	(2)	(3)	(4)	(5)	(6)
Authority	0.174	0.178	0.179	0.172	0.176	0.178
	(0.195)	(0.195)	(0.194)	(0.195)	(0.195)	(0.194)
Transparency	−0.474**	−0.478**	−0.465**	−0.474**	−0.478**	−0.465**
	(0.196)	(0.196)	(0.195)	(0.196)	(0.197)	(0.195)
NoDiscrimination	−0.202	−0.207	−0.197	−0.212	−0.217	−0.206
	(0.198)	(0.198)	(0.197)	(0.200)	(0.200)	(0.199)
Authority*Transparency	0.130	0.141	0.162	0.127	0.137	0.159
	(0.226)	(0.225)	(0.225)	(0.227)	(0.226)	(0.226)
Authority*NoDiscrimination	−0.112	−0.127	−0.141	−0.103	−0.117	−0.133
	(0.224)	(0.224)	(0.223)	(0.227)	(0.227)	(0.226)
Transparency*NoDiscrimination	0.254	0.268	0.251	0.256	0.270	0.253
	(0.227)	(0.228)	(0.226)	(0.227)	(0.228)	(0.226)
Co-players in Favor minus Against	0.334***	0.319***	0.306***			
	(0.048)	(0.049)	(0.048)			
Co-players Sentiment		0.052			0.052	
		(0.049)			(0.049)	
Co-players Policy position			0.103***			0.103***
			(0.037)			(0.037)
Sentiment		−0.027	−0.024		−0.027	−0.023
		(0.062)	(0.062)		(0.062)	(0.062)
Policy position		0.070*	0.057		0.069*	0.056
		(0.041)	(0.042)		(0.042)	(0.042)
Co-players in Favor				0.294***	0.280***	0.268***
				(0.099)	(0.099)	(0.099)
Co-players Against				−0.381***	−0.364***	−0.348***
				(0.111)	(0.111)	(0.111)
Diff Understand AI	−0.033	−0.029	−0.034	−0.034	−0.030	−0.035
	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
Attitudes AI (composite)	0.046	0.040	0.042	0.047	0.041	0.042
	(0.037)	(0.038)	(0.038)	(0.037)	(0.038)	(0.038)
Legal	0.016	0.024	0.034	0.016	0.024	0.035
	(0.334)	(0.334)	(0.336)	(0.334)	(0.334)	(0.335)
Obs.	2,397	2,397	2,397	2,397	2,397	2,397
Controls	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	3646.714	3649.599	3644.350	3648.467	3651.365	3646.141

Note: The table reports ordered logit regressions. Dependent variable is the difference between the opinion about AI before and after chat. Independent variables are the same as described in Table 8. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C6: Opinion about AI described in scenario (Removing participants with incorrect answer to comprehension question)

	(1)	(2)	(3)	(4)
Authority	−0.108	−0.141	−0.241*	−0.232*
	(0.072)	(0.132)	(0.140)	(0.139)
Transparency	0.333***	0.374***	0.467***	0.456***
	(0.072)	(0.129)	(0.134)	(0.135)
NoDiscrimination	−0.050	0.018	0.038	0.025
	(0.072)	(0.127)	(0.133)	(0.134)
Authority*Transparency		0.061	0.037	0.043
		(0.144)	(0.154)	(0.153)
Authority*NoDiscrimination		0.005	0.112	0.095
		(0.144)	(0.154)	(0.153)
Transparency*NoDiscrimination		−0.141	−0.125	−0.095
		(0.144)	(0.153)	(0.153)
Diff Understand AI			−0.081*	−0.069
			(0.044)	(0.046)
Attitudes AI (composite)			0.415***	0.407***
			(0.030)	(0.031)
Biased Judges			0.156***	0.122**
			(0.059)	(0.060)
AI more/less biased			−0.559***	−0.563***
			(0.044)	(0.044)
AI non-discriminatory data			0.348***	0.334***
			(0.037)	(0.037)
Legal			−0.344	−0.357
			(0.261)	(0.267)
Obs.	2,682	2,682	2,664	2,657
Controls	No	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
AIC	7059.931	7064.773	6480.953	6457.497

Note: The table reports ordered logit regressions. Dependent variable is the five-point likert-scale about the opinion of AI as described in the corresponding scenario. Independent variables are the same as described in Table 4. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C7: Opinion change (Removing participants with incorrect answer to comprehension question)

	(1)	(2)	(3)	(4)	(5)	(6)
Authority	0.293	0.303	0.303	0.291	0.301	0.301
	(0.192)	(0.191)	(0.191)	(0.191)	(0.191)	(0.191)
Transparency	−0.415**	−0.412**	−0.402**	−0.417**	−0.414**	−0.404**
	(0.194)	(0.194)	(0.193)	(0.194)	(0.194)	(0.193)
NoDiscrimination	−0.169	−0.177	−0.169	−0.184	−0.192	−0.183
	(0.201)	(0.200)	(0.199)	(0.202)	(0.202)	(0.201)
Authority*Transparency	0.027	0.036	0.055	0.022	0.029	0.049
	(0.225)	(0.224)	(0.224)	(0.226)	(0.225)	(0.225)
Authority*NoDiscrimination	−0.235	−0.256	−0.267	−0.222	−0.242	−0.253
	(0.223)	(0.223)	(0.222)	(0.226)	(0.225)	(0.225)
Transparency*NoDiscrimination	0.374*	0.391*	0.373*	0.378*	0.396*	0.378*
	(0.226)	(0.227)	(0.225)	(0.226)	(0.227)	(0.225)
Co-players in Favor minus Against	0.320***	0.297***	0.291***			
	(0.047)	(0.048)	(0.047)			
Co-players Sentiment		0.080			0.080	
		(0.050)			(0.050)	
Co-players Policy position			0.097***			0.097***
			(0.036)			(0.036)
Sentiment		−0.038	−0.027		−0.037	−0.026
		(0.062)	(0.062)		(0.062)	(0.062)
Policy position		0.105**	0.094**		0.103**	0.093**
		(0.042)	(0.042)		(0.042)	(0.043)
Co-players in Favor				0.249**	0.229**	0.226**
				(0.099)	(0.099)	(0.099)
Co-players Against				−0.401***	−0.375***	−0.366***
				(0.113)	(0.114)	(0.114)
Diff Understand AI	−0.036	−0.030	−0.034	−0.038	−0.032	−0.036
	(0.055)	(0.056)	(0.055)	(0.055)	(0.056)	(0.055)
Attitudes AI (composite)	0.045	0.036	0.037	0.046	0.037	0.037
	(0.036)	(0.037)	(0.037)	(0.036)	(0.037)	(0.037)
Legal	−0.037	−0.033	−0.022	−0.034	−0.032	−0.022
	(0.302)	(0.303)	(0.303)	(0.301)	(0.302)	(0.302)
Obs.	2,473	2,473	2,473	2,473	2,473	2,473
Controls	No	No	Yes	No	No	Yes
Clustered SE	Yes	Yes	Yes	Yes	Yes	Yes
AIC	3709.207	3708.222	3705.308	3710.450	3709.518	3706.661

Note: The table reports ordered logit regressions. Dependent variable is the difference between the opinion about AI before and after chat. Independent variables are the same as described in Table 8. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.