PUBLIC ATTITUDES TOWARD ALGORITHMIC RISK ASSESSMENTS IN COURTS
A DELIBERATION EXPERIMENT

Arna Woemmel
Ph.D. Candidate in Economics
University of Hamburg
arna.woemmel@uni-hamburg.de

Hendrik Huening (University of Hamburg)
Lydia Mechtenberg (University of Hamburg)

Do citizens support the use of predictive algorithms in the criminal justice system?
How important is the ethical design of these tools to them?

Project Overview

2,358 UK citizens discussed a hypothetical scenario of implementing algorithmic risk assessments in UK courts.

- **Groups of 3**, randomly matched.
- Free-form online **messenger chats.**
- Scenario**: Algorithmic risk tool supporting judges** in early-release decisions.
- Elicitation of individual attitudes before and after the chat discussion (**attitude change**).
- Analysis of **chat content**.
- Participant recruitment via Prolific (UK).

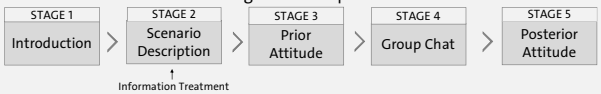
Motivation

- **Procedural Justice:** Perceived fairness and legitimacy as key determinants of compliance with the law (Sunshine and Tyler, 2003; Tyler, 2006; De Cremer and Tyler, 2007).
- **Limitations of Current Literature**: Existing research is limited to survey studies with inconsistent findings (e.g., Simmons, 2017 vs. Wang, 2018).
- **Public Participation in AI**: Increasing calls, both inside and outside academia, for public participation in the AI debate to "strengthen AI democracy" (Acemoglu, 2021), especially through public deliberation (Sætra 2022; Stilgoe, 2024) .
- **Knowledge Gap:** Limited understanding of how people develop beliefs and attitudes about AI (De Freitas, 2023).

Experimental Design

- Participants received a general introduction to algorithms and a scenario describing their use and design features in the criminal justice context.
- Information Treatments in descriptions (2x2x2 between-subjects design, no variation within chat groups):
  - (i) Transparency: High vs. low public oversight;
  - (ii) Fairness vs. Accuracy: Blind vs. non-blind treatment of sensitive attributes;
  - (iii) Judge's Discretion: Full vs. constrained discretion over the algorithm's recommendations.
- Chat: 786 open-ended chat discussions (7 minutes) generating over 15k messages post data cleaning.
- Attitudes measured before and after chat (using Likert-Scales).

Main stages in the experiment:



Data Analysis

- **NLP Methods**: Analysis of discussion flow and trends, focusing on shifts in tone and topics raised over time.
- **OpenAI API 4 model:** Identification of key topics.
- **Correlation Analysis**: Association between topics raised in the chat and attitudes of chat partners with attitude change post-deliberation.

Main Takeaways

- Fragile AI optimism: Stated public support for AI in public institutions may be malleable (caution with one-shot opinion polls).
- People care about fairness, reliability, and individualized determinations.
- Social learning: People are responsive to (online) public discourse, especially when risks are deliberated on in greater depth.

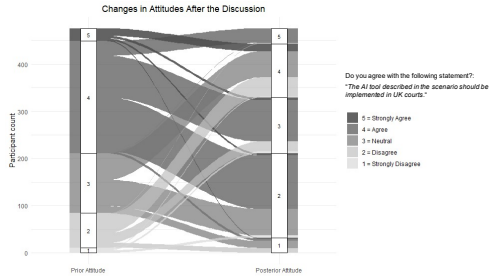
Methods:

- Deliberation study + NLP + LLM: Promising approach for studying people's reasoning, attitude formation, and attitude strength.

Results

Significant Changes in Attitudes

- Prior attitudes: Overall positive, especially with high public oversight (p=0.021) and full judge's discretion over the algorithm (p=0.092).
- Posterior attitudes: Deliberation led to **negative attitude shifts** (p<0.001).



- 26% of initial supporters became more skeptical; no such upward shift for initial opponents.
- Convergence of attitudes: Deliberation diminished treatment effects.
- Neutral sentiment: Discussions were characterized by neutral, non-emotional language (possible reasons: unfamiliarity with the topic; topic not (yet) politically charged?).

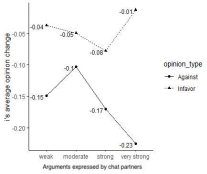
Key Topics (by frequency)

1. *"judge bias"*, *"algorithmic bias*"*, *"cost savings"*
2. *"support tool"*, *"human element"*, *"public trust"*
3. *"less noise/consistency"*, *"lack of individualized treatment*"*, *"technical errors*"*
4. ..

Balanced distribution of positive and negative arguments.
*(*correlation with attitude change)*

Negativity Bias

1. Attitude changes more sensitive to counterarguments.
2. Presence of a single opponent in chat group associated with stronger attitude change (p<0.001).
3. Greater topic heterogeneity in opposing arguments than in supportive ones.



References:
- Acemoglu, D. (Ed.). (2021). Redesigning AI. MIT Press.
- 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.
- De Freitas, J., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, 7(11), 1845-1854.
- Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, 30(3), 291-323.
- Sætra, H. S., Borgebund, H., & Coeckelbergh, M. (2022). Avoid diluting democracy by algorithms. *Nature Machine Intelligence*, 4(10), 804-806.
- Simmons, R. (2017). Big data and procedural justice: Legitimizing algorithms in the criminal justice system. *Ohio St. J. Crim. L.*, 15, 573.
- Stilgoe, J. (2024). AI has a democracy problem. Citizens' assemblies can help. *Science*, 385(6711), eadr6713.
- Sunshine, J., & Tyler, T. R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. Law & Society Rev., 37(3), 513-547.
- Tyler, T. R. (2006). Why people obey the law. Princeton University Press.
- Wang, A. J. (2018). Procedural justice and risk-assessment algorithms. Available at *SSRN 3170136*.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
European Commission
Horizon 2020
European Union funding for Research & Innovation