On Human Factors in Machine Fairness: Essays in Behavioral Economics

Universität Hamburg
Fakultät Wirtschafts- und Sozialwissenschaften

Kumulative Dissertation

Zur Erlangung der Würde eines Doktors
der Wirtschafts- und Sozialwissenschaften
„Dr. rer. pol."
(gemäß der PromO vom 18. Januar 2017)

vorgelegt von

Arna Carolin Wömmel
aus Mettingen, Deutschland

Hamburg, 23. Juni 2025

Vorsitzender: Prof. Dr. Andreas Lange

Erstgutachter: Prof. Dr. Gerd Mühlheußer

Zweitgutachterin: Prof. Dr. Judith Simon

Datum der Disputation: 29.07.2025

Acknowledgements

First and foremost, I would like to thank my first advisor, Gerd Mühlheußer, for his unwavering support over the past years. Gerd has made me a better thinker, writer, and researcher. He gave me the freedom to pursue my own ideas, explore new methods, and learn from my mistakes. He also gave me opportunities to work on topics I truly care about, to present my projects early on, and to spend time abroad. His honesty and reliability have been invaluable in writing this thesis. Thank you, Gerd.

I am also very grateful to my second advisor, Judith Simon, for her constructive feedback and thoughtful guidance on my projects, and beyond. Thank you, Judith, for taking the time to comment on drafts, attend my talks, and check in with me. This thesis greatly benefited from both Gerd's and Judith's openness to interdisciplinary supervision and projects across fields.

Thank you to my co-authors, Hendrik Hüning and Lydia Mechtenberg, for their productive collaboration and essential contributions in initiating our project. I'm especially grateful to Hendrik for his guidance as I began working on this thesis.

I was fortunate to be part of in the DFG Graduate Program *Collective Decision-Making*, which provided excellent conditions for writing this thesis. I thank all PIs, fellow researchers, and everyone involved for making this experience so enriching. A *huge* thank you goes to Anke Gerber for her outstanding work in developing and managing the program.

I am very grateful for the generous financial support that made the research in this thesis possible. I would like to thank the German Research Foundation (DFG), the European Unions Horizon 2020 research and innovation program, Gerd's chair, and the German Institute for Economic Research (DIW Berlin) for their funding. Special thanks go to Martin Gerike at DIW for his outstanding coordination of the SOEP-IS module used in Chapter 4.

All projects greatly benefited from formative research stays abroad. I was fortunate to spend time at Columbia University's Data Science Institute, Columbia Law School, and ETH Zurich. I am very grateful to the DFG and the graduate program for making these visits possible and to my hosts - Eric Talley, Talia Gillis, and Alexander Stremitzer - for their generous hospitality. Many thanks also to those whose kindness and openness made me feel welcome there from the very beginning: Luca Baltensperger, Raquel De Haro, Janka Dehli, Zuko Garagic, Jakob Merane, Tamalika Mukherjee, Alessandro Tacconelli, Claudia Shi, and Keyon Vafa.

Aileen Nielsen provided valuable guidance and perspective throughout. I learned a lot from her.

My social network at UHH has been a tremendous help these past few years. I am especially grateful for the close friendship of Julia Jakobi, Juliane Koch and Jasmin König. They made Ph.D. life joyful and fun. I also warmly thank the Koch family for so often letting us stay at their beautiful holiday home by the Baltic Sea; the cold swims and cozy dinners there brought us rest and renewal.

Susanne Wilmer, my high school teacher, holds a special place in my acknowledgments. She inspired me to study economics and remains, without doubt, the best teacher I've ever had (and I've had quite a few). Her example continues to shape the way I think about and practice teaching.

My friends outside academia have meant a lot to me throughout this time. Thank you for keeping me grounded and for showing interest in my research. There are three special people who have carried me through many chapters of life, including the writing of this thesis: Adriane, Julie, and Nora. They have been my constant source of support, laughter, and comfort, since childhood. Words cannot express how grateful I am for our friendship.

Lastly, I would like to thank my family. I am very grateful to my parents, Barbara and Ekkehard Wömmel, who invested all kinds of resources and made many sacrifices for my upbringing and education. They always encouraged me to explore my interests and find my own path, even when it led me far from familiar ground.

I am also grateful to my sister Kristin and her family, Hermann, Finn, and Frida, for the joy they bring, their steady presence, and the many wonderful respites they offered.

I will never forget the support I received from my sister, Inga, and her family, Sven and Klara, while writing this thesis. From the very beginning, they made me feel at home in Hamburg, surrounding me with unwavering generosity, encouragement, and a deep sense of peace. I will hold close the many special moments we shared during this time. Forever thank you.

Contents

Chapter 1.

Introduction

Digital technologies and artificial intelligence (AI) have now permeated nearly all areas of public and private life. Driven by rapid advances in machine learning and the availability of large data, these systems have become central to high-stakes decision-making processes and are altering how individuals work, communicate, and access information. Given their far-reaching societal implications, a growing body of research and policy debate has focused on ensuring that these technologies are developed and deployed in ways that are both ethically sound and socially beneficial.

As a result, a range of technical and regulatory principles has been implemented to guide their design, implementation, and governance (Barocas et al., 2023). This often involves a delicate balancing act. For instance, when training and developing such systems, attention is given to which aspects of society they should reflect, such as social values, and which they should explicitly not reflect, such as social injustices and discrimination (Barocas and Selbst, 2016; Gillis and Spiess, 2019; Simon et al., 2020). Regarding the technological transformation of the labor market, there is a trade-off between leveraging the benefits of technological advancement and ensuring that automation and unequal access do not exacerbate social disparities (Acemoglu, 2021a). Finally, an increasingly central question is *who* should define these principles. A growing number of scholars argue that broader public involvement is needed to enhance democratic legitimacy and reduce the concentration of power among private actors (Acemoglu, 2021b; Sætra et al., 2022).

In this thesis, I draw on methods and literature from behavioral economics to empirically demonstrate that, although well-intentioned, these efforts may be undermined by human behavior. These principles are often developed under the assumption that these technologies operate either in fully automated settings without human involvement or are used by agents who are rational and unbiased. I show that these assumptions may not hold, leading to unintended consequences. First, individuals often exhibit preference misrepresentation and fail to articulate attitudes that reflect their true preferences. Second, even when algorithmic tools are optimized for accuracy and fairness, these system-level gains can break down when used by decision-makers whose behavior deviates from payoff maximizing rational behavior (which would predict full adoption in this case), and distorting biased beliefs and tastes. The results in this thesis also challenge the growing assumption that such technologies will reduce inequality in the labor market (Autor et al., 2024; Brynjolfsson et al., 2025). While they may increase productivity and reach across occupational boundaries, the results in this thesis suggest that adoption may be unevennot only due to skill disparities as shown in other studies (Humlum and Vestergaard, 2025), but also because of biases in individuals beliefs about their own ability to leverage these tools. These findings should not be interpreted as a critique of efforts to promote the ethical design and governance of new technologies. Rather, they highlight that human behavior often differs from the assumptions underlying much of the current discourse. For this reason, behavioral mechanisms should be given greater consideration when designing and evaluating institutional frameworks for emerging technologies, to ensure that the intended objectives are realized in practice.

Chapter 2[1], titled ”Fragile AI Optimism” (coauthored with Hendrik Hüning and Lydia Mechtenberg), addresses the issue of the *public legitimacy* in the use of predictive machine learning tools in public decision-making. For institutions to maintain public trust and ensure compliance, broad public approval is essential. In line with this, a growing body of scholarship has called for greater public involvement in debates surrounding AI and advocates for more inclusive governance frameworks to ensure that these technologies reflect diverse societal values and uphold democratic principles. An increasing number of surveys and opinion polls have sought to elicit public perceptions of emerging technologies, including which features and institutional safeguards are prioritized (Starke et al., 2022). However, existing findings tend to be inconsistent, highly context-dependent, and sensitive to survey design and framing (Chen et al., 2022; Nussberger et al., 2022; Zhang and Dafoe, 2019; Schmager et al., 2024). Overall, there is a growing tendency indicating that the public is generally open to the use of predictive machine learning tools, including AI, for supporting in the public sector, provided certain safeguards and normative principles are guaranteed (Dietvorst et al., 2018; Kennedy et al., 2022; Chen et al., 2025).

The contribution of this chapter is to provide empirical evidence that such stated public approval is fragile. It presents results from an online deliberation study with 2,358 UK citizens focused on the use of predictive machine learning tools in the criminal justice system. In the study, the participants state their approval of these systems both before and after discussing it with two other participants in online messenger chats. The data shows that while initial approval aligns with prior survey findings, especially when certain normative criteria are met (Gesk and Leyer, 2022; Kennedy et al., 2022; Scurich and Krauss, 2020; Sidhu et al., 2024), the supportive attitudes converge downward after the discussion, and the impact of specific design features diminishes. In contrast, participants who entered the discussion with critical views are especially resistant to change. The results of the quantitative analysis of chat transcripts

[1] An earlier paper of this project was recognized as *Best Paper* in the category ”Technology, Information, and Privacy” at the 2023 American Law and Economics Association Conference in Boston, and awarded the “Theodore Eisenberg Prize” at the 2024 Conference on Empirical Legal Studies in Atlanta.

suggests that these individuals contribute more frequently, raise a broader range of arguments, and are less responsive to counterarguments.

Following the behavioral economics literature, we interpret these findings as indicator that stated public approval reflects higher uncertainty. Stated approval for these tools appears to be quickly increased through favorable framing, but just as quickly eroded through deeper deliberation. This asymmetry in attitude strength - weak approval yet robust resistance - has several implications. First, survey-based measures of public approval, including those showing that certain tool or institutional features increase public support, should be interpreted with caution. Stated approval may reflect uncertainty rather than stable preferences, especially given the novelty and complexity of the domain. Second, there is a need to educate and inform the public about the technological transformation of public institutions. Ensuring meaningful public input and evaluation requires that citizens are equipped to form robust, well-informed views. The findings here indicate that there is still an inherent uncertainty about this topic. Third, the robustness of opposition suggests that in case of well-designed and socially beneficial systems, it may require targeted efforts to build public trust and to achieve the intended (welfare) gains. Finally, our data is consistent with prior research, in showing that individuals value human decision authority and express concern about algorithmic bias, which was a key driver of opposition and negative attitude shifts following deliberation.

In Chapter 3, titled "Algorithmic Fairness and Human Discrimination", I show that these normative preferences for non-discriminatory algorithmic design may fail to hold when individuals shift perspectives: from evaluating such tools as citizens to applying them as decision-makers.[2] Using an online lab-style hiring experiment, I find suggestive evidence that algorithmic tools that provide performance predictions of job candidates are less adopted when they explicitly exclude sensitive attributes from their training data. Participants first show that they are more conservative in their belief updating about candidates when presented with an algorithm that is blinded to protected group membership. Then, regarding subsequent hiring decisions, I find that discrimination under this hiring significantly increase under this algorithm. In short, the system-level intervention may reduce adoption, meaning decision-makers do not fully update according to the algorithm's signals, and statistical discrimination may persist. The data shows that this is not driven by an intervention into the training data per se, but pertains specifically to the exclusion of sensitive attributes. Moreover, following the economics literature on discrimination, I additionally find that human decision-makers base their decisions not only on their beliefs about performance differences, but also on personal preferences for individuals from cer-

[2] Similarly, in the context of autonomous vehicles, Bonnefon et al. (2016) find that while people state that they endorse utilitarian AVs (vehicles that would sacrifice their passengers for the greater good) in principle and want others to adopt them, they would not be willing to purchase such a vehicle.

tain groups ("taste-based discrimination"). In such cases, neither the algorithm nor its fairness features are sufficient to prevent discrimination. The algorithms objective functionmaximizing predictive accuracy simply diverges from the human decision-makers utility function, involving a preference for individuals from certain groups.

These findings add a behavioral perspective to the established debate on *disparate treatment* and *disparate impact* in algorithmic decision-making. The existing discussion has mainly focused on the legal prohibition of using sensitive attributes in algorithmic models, which often stands in tension with technical requirements for accuracy and fairness (Hellman, 2020). This thesis shows that excluding sensitive attributes can also produce unintended behavioral effects. Individuals may become less willing to adopt algorithmic tools when these tools are subject to certain fairness interventions. This reluctance highlights a potential trade-off: while fairness interventions aim to reduce discriminatory inputs, they may also reduce acceptance and limit the effectiveness of such systems (if designed well) in practice. From a regulatory standpoint, this suggests that two foundational principles of AI governance, i.e., non-discriminatory design and human oversight, may in combination produce unintended consequences unless behavioral mechanisms are explicitly accounted for.

Chapter 4, titled ”Social Disparities in Digital Skills: Evidence from Germany”, relates to the longstanding issue of the *digital divide.* The chapter shows that the digital transformation of the labor market may exacerbate existing inequalities by amplifying unequal preparedness for the growing demand for digital skills. Using data from a representative German household panel, the chapter documents persistent gender and socioeconomic disparities in job-relevant digital skills. Men and individuals with higher levels of education perform better on all skills measures. A novel contribution of this study is that these two groups also show greater confidence in outperforming others, even when actual skills levels are held constant. These belief gaps appear not to be driven by overconfidence (which is in contrast to much of the prior behavioral economics literature, e.g.,Malmendier and Tate (2005)), but by underconfidence among highly skilled women and individuals with lower education levels. This may be important for the discussion on the digital divide and growing labor market inequalities in the face of the technological transformation: The behavioral economics literature shows that such beliefs about ones own ability are a key determinant for individuals’ labor market decisions, including application decisions, wage negotiations, and investment in further training and education. When both, actual digital skills and confidence in these skills, are unequally distributed, such dynamics can reinforce each other and contribute to the persistence of inequality in accessing and leveraging emerging job market opportunities. At the same time, this chapter provides suggestive evidence that intergenerational socioeconomic disadvantage may play a reduced role in determining labor market outcomes (Corak, 2013). I find no significant relationship between early-life socioeconomic status and digital skills or confidence in these skills in adulthood. While digitalization may introduce new forms of exclusion through skill and belief disparities, it may also weaken the link to inherited disadvantage. Still, digital skills are strongly associated with education, which is typically strongly associated with socioeconomic background, but requires a distinction.

These findings highlight three key policy implications for the digital labor market. First, addressing gaps in digital skill proficiency alone may prove insufficient; policymakers must also target confidence gaps, particularly among women and individuals from less advantaged educational backgrounds. Interventions should combine skills training with programs that foster encouragement for participation in an increasingly digitalized work environment. Second, education remains a key mechanism for equalizing opportunity, also in the digital economy. To be effective, policies should embed digital training early in the education system, where disparities in access and engagement are smaller and interventions are more likely to have broader and lasting effects. Finally, the weakening association between early-life socioeconomic status and adult digital preparedness suggests a new channel for social mobility. This also implies that targeted interventions, especially those anticipating and addressing future skill demands, may effectively reduce persistent inequalities and expand labor market opportunities across socioeconomic groups.

Chapter 2.

Fragile AI Optimism[1]

Abstract

We study how public attitudes toward AI form and shift using an online deliberation study with 2,358 UK citizens in the criminal justice context. First, we replicate prior survey evidence suggesting widespread public support for adopting AI as a decision-support tool, particularly when certain fairness features are met. We then show that this stated support is fragile: it declines significantly with group deliberation, as supporters are 2.6 times more likely than opponents to change their attitudes. Results from the quantitative text analysis of the chat content indicates that opponents contribute more arguments, both in terms of frequency and topic range, and supporters are more responsive to counterarguments. These results suggest that stated public support for AI reflects lower attitude strength as it appears to be quickly raised through informational framing but just as quickly reversed through deliberation. More broadly, they caution against inferring public legitimacy of increased AI deployment from stated support alone.

Keywords: Deliberation, Public Attitudes toward AI, Algorithmic Risk Tools, AI and Law, Text Analysis

JEL Classification: C91; J71; O33

2.1 Introduction

Public approval is a key requirement for institutional legitimacy and compliance (Tyler and Huo, 2002; Sunshine and Tyler, 2003; Tyler, 2006). The rapid and widespread integration of new technologies, such as artificial intelligence (AI), in the public sector has consequently prompted efforts to elicit public attitudes toward these tools and their deployment (Acemoglu, 2021b; Sætra et al., 2022; Stilgoe, 2024). Yet public legitimacy, to be effective and enduring, requires that public attitudes be well-informed and firmly held, not merely transient (Suchman, 1995; Fishkin, 2018; Lafont, 2023). How informed and robust public attitudes toward AI are is an intricate and open question: the topic is rapidly evolving, media coverage is polarized, and stances are not (yet) anchored in political identities.

In this chapter, we provide both causal and descriptive evidence from a collective deliberation experiment showing that support for AI is less stable than opposition to it. Although many participants initially express support, this stance proves more malleable and appears to reflect weaker underlying attitudes. We demonstrate this asymmetry in attitude strength based on three complementary indicators: first, supporters are significantly more likely than opponents to change their views following deliberation; second, they are more responsive to

[1] This chapter is co-authored by Hendrik Hüning and Lydia Mechtenberg.

counterarguments; and third, opponents provide stronger reasoning for their stance in terms of both frequency and heterogeneity.

The deliberation study was conducted online and involved 2,358 UK citizens deliberating on the implementation of AI in the UK criminal justice system. Participants were presented with a scenario in which AI served as a decision support tool for judges, providing predictions about future criminal behavior. The scenario varied along three dimensions involving a 2 x 2 x 2 between-subjects design: (i) the degree of judicial discretion, (ii) the level of public oversight, and (iii) whether sensitive attributes were excluded at the cost of predictive accuracy. Each participant was assigned to one scenario, defined by a combination of the three dimensions, and placed in a group with two others who received the same version to discuss in a real-time, text-based messenger chat. Participants stated their individual attitudes before and after the chat. We measure the impact of deliberation on attitude change and apply quantitative text analysis to the chat transcripts to explore how these attitudes form and shift. Specifically, we classify message content by stance using supervised machine learning methods[2] and examine how message stance is associated with participants (i) initial attitudes and (ii) subsequent attitude change. We train the classifiers on labeled data collected during the experiment, where participants provided separate free-text responses outlining the perceived advantages and disadvantages of introducing such AI tools in the criminal justice system.

We report three main findings. We first replicate prior survey results showing that stated attitudes toward AI decision-support tools are overall favorable (54% favorable, 29% opposed, and 14% neutral), but sensitive to design. Approval significantly increases when tools are under public oversight and when human discretion is retained, while attitudes remain ambivalent when deciding between the two fairness criteria of higher predictive accuracy versus non-discriminatory input factors. The first key finding of this study is that deliberation significantly reduces both overall approval and the heterogeneity raised through the framing in attitudes. Regardless of the information participants initially received about the AI, their views became more skeptical and more uniform following deliberation. The second key finding is that changes in attitudes are driven by initially confident supporters: those who were firmly in support of AI before the discussion were 2.6 times more likely to shift their views than those who initially strongly opposed it. This is notable considering that most participants were matched with other participants holding favorable attitudes. The third key finding provides causal evidence that opponents were more persuasive in the chat. Conditional on prior attitude, encountering at least one opponent in the chat increased the odds of attitude change by approximately 64.2% compared to encountering at least one supporter. Descriptive results from the text analysis

[2] We use BERT, Lasso, and Wordscore (Bag-of-Words approach) for this task and cross-validate the output across these three methods.

strengthens this by showing that opponents were more active in the chat, raising more negative arguments in terms of both frequency and topic range. Opponents were also less susceptible to counterarguments than supporters.

These results suggest that stated support for AI may be quickly generated through informational framing but just as quickly eliminated through deliberation. We interpret such reversals as indicative of weakly held attitudes and underlying uncertainty. According to economic models of social learning, individuals with weak priors are more likely to update their beliefs in response to social cues, i.e., others actions or information, which results in a convergence in opinions and behavior (Banerjee, 1992; Avery and Zemsky, 1998; Bikhchandani et al., 2024). This interpretation is also consistent with models of Bayesian persuasion, which show that individuals with weak or diffuse priors are more responsive to external information, strategic framing, and counterarguments, even when signals are noisy or limited (DellaVigna and Gentzkow, 2010; Kamenica and Gentzkow, 2011). A sender has greater influence over a receivers opinions or actions when the receivers prior beliefs are weak. We therefore interpret our findings in reverse: individuals who are more responsive to both (i) informational framing in the experimental prompt, and (ii) social cues during deliberation likely hold weaker priors and greater uncertainty. Conversely, regarding the sender behavior, the economic literature shows that those with more informed priors and more certain preferences are more likely to express their views when communicating and deliberating with others (Schwardmann et al., 2022). They individuals have been shown to act as more effective advocates and persuaders (DellaVigna and Gentzkow, 2010; Kamenica, 2019; Fafchamps et al., 2024).

These findings contribute to understanding the mixed empirical evidence on public attitudes toward AI, with some studies reporting support and others reporting opposition (we replicate the former). Prior work has attributed such variation to high context sensitivity, emphasizing the importance of specific features of the AI system (e.g., accuracy, fairness constraints), characteristics of the respondent (e.g., demographic traits, AI familiarity), and the nature of the alternative (e.g., a human decision-maker, human worker) (Dietvorst et al., 2015; Castelo et al., 2019; Logg et al., 2019; Starke et al., 2022). We offer a complementary explanation: given the novelty and complexity of AI systems, individuals often face inherent uncertainty about how to evaluate them. As a result, they may be especially susceptible to persuasive framing and express approval in response to optimistic or authoritative descriptions. However, such stated support may be fragile, easily reversed when exposed to countervailing information, and does not necessarily reflect well-informed and robust preferences. This interpretation aligns with broader concerns in the economics literature about the limited validity of stated responses in one-shot surveys (Bertrand and Mullainathan, 2001; Hausman, 2012; Cohen et al., 2020). Our findings highlight the importance of more robust methods for eliciting public AI attitudes, such as open-ended responses, deliberative formats, and repeated measurement over time (Ferrario and Stantcheva, 2022).

Methodologically, we make two main contributions. First, we add to the literature on free-form communication experiments (Brandts et al., 2019; Muehlheusser et al., 2024; Promann et al., 2025), particularly those involving messenger-based chat interventions (Hüning et al., 2022b; Corgnet et al., 2024; Grunewald et al., 2024; Hausladen et al., 2024). This setup mirrors the digital environments in which people frequently exchange information and opinionssuch as social media and messaging platformswhile also ensuring anonymity and minimal infrastructure. It allows us to collect rich communication data to study first-order concerns and the formation of opinions through peer interaction. We also contribute to the emerging literature in economics that uses free-form, open-ended responses to elicit beliefs and opinions (Ferrario and Stantcheva, 2022; Haaland et al., 2025). Second, we add to the growing application of quantitative text analysis methods in economics (Gentzkow et al., 2019; Ash and Hansen, 2023), particularly recent approaches that analyze unstructured communication data using machine learning (Hüning et al., 2022a; Lange et al., 2022; Andres et al., 2023; Ash et al., 2025). Our approach is novel in that it derives labeled training data directly from participants' free-form text responses. We also deviate from prior literature by validating the classification output using multiple classifiers, which offers a scalable, efficient, and less error prone alternative to manual human annotation.

This chapter proceeds as follows. Section 2 describes the experiment and data. Section 3 presents the main results. Section 4 discusses the findings, and Section 5 concludes.

2.2 Experiment

This section describes the experimental design (Section 2.2.1) and the sample (Section 2.2.2).

2.2.1 Experimental design

Figure 2.1 outlines the stages of the experiment. Full instructions are provided in Appendix 2.A.1. At the beginning of the experiment, participants report their level of knowledge of AI. [3] Participants also report how familiar they think the general population is with AI as a relative measure of their own AI knowledge. They then answer two questions about their general attitudes toward AI: one about the strength of their positive views and one about the strength of their negative views. This approach allows us to identify ambivalent participants and avoid

[3] Throughout the experiment, we use the term AI rather than algorithm or "algorithmic risk tool" because it is more familiar to the general public.

conflating neutral responses with a mix of strong opposing views. All attitudes are elicited using 5-point Likert scales.

Participants then read a brief explanation of how AI tools would be used in the criminal justice system. The scenario describes a hypothetical situation in which judges receive AI predictions about a prisoner's likelihood of reoffending to inform decisions about early release. We chose this setting of early-release decisions because it is more familiar and relevant to UK participants, unlike bail decisions, which are more prominent in the U.S. context. The explanation emphasizes that the AI serves as an advisory tool and that judges retain final decision-making authority. It also notes potential benefits such as saving time and reducing public spending. To ensure comprehension, participants answer a multiple-choice question about the tool; those who respond incorrectly receive a clarifying explanation before proceeding.

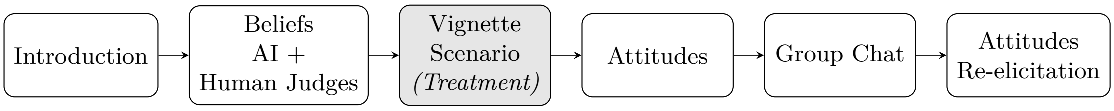


Figure 2.1: Overview of Stages in the Experiment

Next, participants are introduced to the issue of bias in human and algorithmic decision-making. They learn that algorithmic predictions can reflect statistical patterns in historical data, including biased or discriminatory decisions. They are also informed that human judges may be biased, and that evidence shows judicial decisions can be influenced by stereotypes. Finally, they are explained that some AI tools exclude sensitive variables from training data to meet fairness criteria, which may reduce predictive accuracy.

We then elicit participants beliefs about these issues. First, they are asked how biased they perceive the average judge to be. Second, they indicate whether they believe AI systems trained on all available data produce more or less biased predictions than the average judge. Third, they are asked whether they believe an AI that excludes sensitive features, such as race, is less accurate than one that includes them. All responses are measured on 5-point Likert scales.

As a next step, participants are introduced to the specific scenario they are asked to evaluate. The scenario describes an AI tool that assists judges in making early release decisions. It explains that the tool generates categorical risk scores (high, medium, or low) that reflect estimates of a prisoner's likelihood of reoffending if granted early release, similar to tools already used to support sentencing and release decisions in other countries. It makes clear that the AI serves solely as a support tool, not a replacement for judicial decision-making. The scenario varies across participants along three factors, each with two levels, resulting in a 2Œ2Œ2 between-subjects design: (i) whether the judge retains full discretion or is required to incorporate the tools prediction; (ii) whether the tool is developed and monitored by a public institution or a private company; and (iii) whether sensitive attributes are excluded from the training data, potentially at the cost of reduced predictive accuracy. An overview of the factor variation is presented below in Table 2.1.

Table 2.1: Factor Variation in Scenario Descriptions

Factor	**(1)**	**(0)**
Restricted Inputs	Sensitive characteristics (e.g., race) excluded from training data to meet fairness standards; may reduce prediction accuracy.	All available data used in training, including sensitive characteristics, to maximize prediction accuracy.
Judge Full Discretion	Judge retains full discretion; AI tool is used only as optional input.	Judge must incorporate the AI prediction into the final decision.
Public Oversight	AI developed by a public institution with public access and oversight.	AI developed by a private company without public access or oversight.

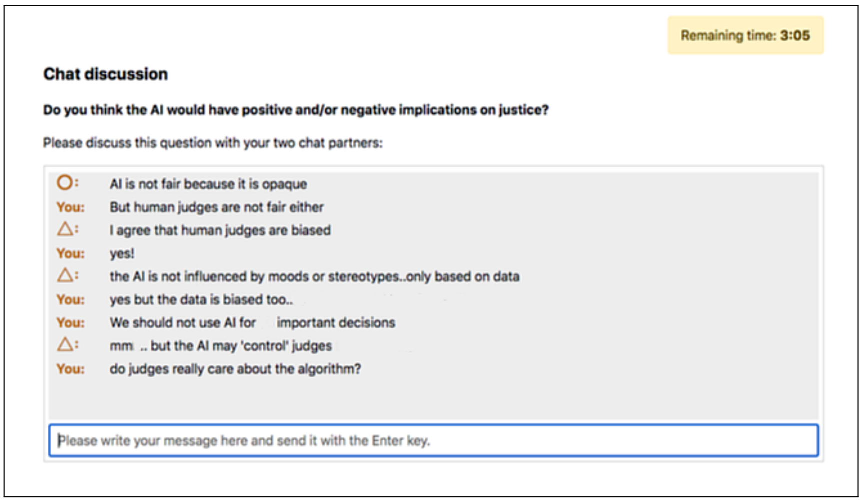
Notes: Factor variations for scenario descriptions. Participants were randomly assigned to one level of each of the three factors, resulting in a $2 \times 2 \times 2$ between-subjects design. Treatment level (1) indicates feature presence; level (0) indicates absence or an alternative.

After reading the scenario, participants state their level of approval for implementing the described AI tool in the UK criminal justice system. They are reminded that there are no right or wrong answers and that their considered personal judgment is what matters most. While stated responses are not monetarily incentivized (doing so would be conceptually and methodologically problematic), participants are told that findings from studies like this one are often used to inform real-world policymaking. Attitudes are measured on a 5-point Likert scale.[4] In addition, participants respond to two open-ended questions about the perceived benefits and drawbacks of implementing the tool.[5] Implementing these free-text responses offers two main advantages. First, by prompting participants to articulate both potential benefits and drawbacks before entering group deliberation, the design introduces a brief phase of individual deliberation. Comparing these responses with those later expressed in the group discussion allows us to isolate the specific role of persuasion, i.e., to distinguish the effects of collective deliberation from individual reasoning by examining how participants' arguments evolve in response to others' contributions. Second, these individual responses are used to fine-tune machine learning models for analyzing the chat transcripts, thereby improving the accuracy of stance classification in the chat content.

[4] The exact wording is: How strongly do you agree with this statement?: The AI, as explained in this scenario, should be used in UK criminal courts.

[5] The exact phrasing of the two questions: Would this AI, as described in the scenario, have positive [negative] effects? If yes, why? Before filling in the two text boxes, participants are reminded that we are interested in their personal judgment and that there are no correct or incorrect answers.

Each participant is then introduced to a real-time group chat session and matched with two other participants who received the same specific scenario description. [6] Prior to joining the group discussion, participants are informed that all communication will remain anonymous. Each chat session is scheduled to last exactly seven minutes and concludes automatically when the time limit is reached. Figure 2.2 shows the chat interface used in the experiment. To ensure anonymity and avoid any social associations, the two other participants are represented only by geometric symbols, a triangle and a circle.



Remaining time: **3:05**

Chat discussion

Do you think the AI would have positive and/or negative implications on justice?

Please discuss this question with your two chat partners:

○: AI is not fair because it is opaque
You: But human judges are not fair either
△: I agree that human judges are biased
You: yes!
△: the AI is not influenced by moods or stereotypes..only based on data
You: yes but the data is biased too..
You: We should not use AI for important decisions
△: mm .. but the AI may 'control' judges
You: do judges really care about the algorithm?

Please write your message here and send it with the Enter key.

Figure 2.2: Chat Interface

Note: This figure shows the chat interface presented to all participants. The messages are fictional and originate from test runs; they do not reflect actual participant data. There was no monitoring of the chats.

After the chat, participants are shown their initial response on the Likert scale and asked whether they wish to revise it. If they choose to do so, they can update their response using the same 5-point Likert scale. All participants are then presented with a free-form text box asking them to explain why they did or did not change their attitude. Finally, participants rate how persuasive they found their chat partners using a 5-point Likert scale. The study concludes with the collection of demographic information.

The study was fielded in October and November 2021, with participants recruited through the Prolific platform. Each of the eight treatment groups was conducted on separate days (see Appendix Table A1 for exact session dates). We conducted a pilot study with 25 students in summer 2021 at the University of Hamburgs online laboratory to confirm the technical feasibility. The fully computerized experiment was implemented using oTree (Chen et al., 2016).

[6] In cases where matching is not possible due to technical issues or participant dropout, individuals are redirected to the end of the study.

During our initial eight sessions, a sampling error on the Prolific platform led to an oversampling of female participants.[7] To address this imbalance, we conducted additional re-sampling sessions with a focus on increasing the proportion of male participants.

2.2.2 Sample

Overall, we recruited 2,864 participants from the UK general population exclusively through the Prolific platform. Prior to the chat phase, the dropout rate was 5.4%. Participants who could not be assigned to a chat group due to real-time matching constraints were automatically excluded. We also exclude from the final sample those matched with only one other person to ensure consistent chat conditions. The final sample consists of 2,358 participants (82% of those recruited) and 786 chat groups. From this sample, we collected a total of 15,429 chat messages.

Of the final sample, 68% are female, 31.4% male, and 0.6% identify as other. Individuals with a college degree are moderately overrepresented compared to the general UK population. Detailed summary statistics are provided in Appendix Tables A4 through A6.

Regarding participants knowledge and beliefs (elicited after the introduction to potential biases in both human judges and AI systems, and before the scenario), participants self-report a moderate level of general AI knowledge (mean=3.14, SD=0.97 on a 1–5 scale), comparable to that of others in the UK population (mean difference=–0.08, SD=0.93). They perceive human judges as moderately biased (mean = 3.04, SD=0.72). Their beliefs about AI are overall optimistic. On average, participants tend to believe that AI predictions with restricted inputs remain accurate (mean=3.29, SD=1.07), though responses to this item exhibit greater variance. They also consider AI to be less biased than human judges (mean=2.61, SD=0.99, values below 3 indicate favorability toward AI).

2.3 Results

This section reports the main findings. Section 2.3.1 describes attitudes prior to group discussion and the effects of the information treatments. Section 2.3.2 presents attitudes after the group chat. Section 2.3.3 turns to causal determinants of attitude change (group composition). Section 2.3.4 presents descriptive results from the chat content.

2.3.1 Prior attitudes

After reading the scenario, the majority of participants report favorable attitudes. Across treatment conditions, 51–57% express approval or strong approval of implementing the AI according to the scenario ($mean = 3.25$, $SD = 1.01$). Disapproval (*disagree* or *strongly disagree*) is less fre-

[7] This issue affected multiple studies during that period and was not specific to our experiment.

quent and ranges from 24–33% across treatments. Neutral responses, representing the midpoint of the scale, are the least common, with 15–20% of replies (for detailed results see in Appendix Table A5). This finding of widespread stated public approval of data-driven prediction tools in decision-making replicates several other survey studies (e.g., Logg et al. 2019), including those focusing on the public sector and criminal justice(Gesk and Leyer, 2022; Kennedy et al., 2022; Scurich and Krauss, 2020; Sidhu et al., 2024).

The data further confirm prior survey research showing that public attitudes are sensitive to specific design features of algorithmic tools (e.g., Jussupow et al. 2020; Nussberger et al. 2022; Starke et al. 2022). As shown in Figure 2.3, we replicate previous findings that approval is significantly higher when the tool is developed by a public institution rather than a private company ($p < 0.001$)Zhang and Dafoe 2020; Kennedy et al. 2022) and when humans retain full discretion over the tool, in particular judges in the criminal justice settings ($p = 0.098$; Bogert et al. 2021; Alon-Barkat and Busuioc 2023; Simmons 2017; Kennedy et al. 2022; Chen et al. 2022, 2025). Finally, in the more complex treatment condition, i.e., where participants must weigh the trade-off between excluding sensitive variables and improving accuracy, we find no significant effect ($p = 0.8128$). This null result is consistent with mixed findings in prior research on public approval of different fairness metrics (e.g., Saxena et al. 2019, 2020; Nussberger et al. 2022; Starke et al. 2022; Wang et al. 2022; Bansak and Paulson 2024). Our data thus corroborate these prior findings that the public values both accuracy and fairness in predictive tools, but often exhibit ambivalence or uncertainty when these objectives are in tension.

Appendix Table A7 presents corresponding results from ordered logistic regressions including the associations between AI knowledge, beliefs about AI and judge bias, demographics, and support for AI in court decisions. Beliefs that judges are biased, that AI is less biased than judges, and that AI remains accurate when restricted are positively associated with attitudes. Among demographic factors, only age is associated with attitudes, which confirms prior research showing that acceptance of AI declines with age (Kelly et al., 2023; Kim and Peng, 2024). With respect to our experimental design, which involved endogenous participant matching for the chat discussions, the fact that a majority of participants initially stated favorable attitudes implies a higher likelihood of encountering supporters during deliberation.

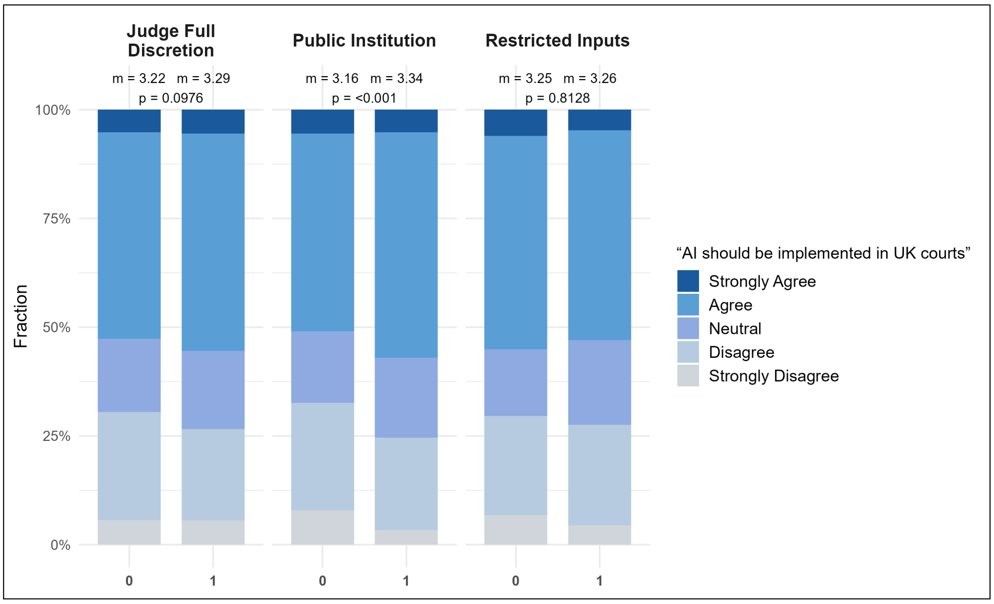


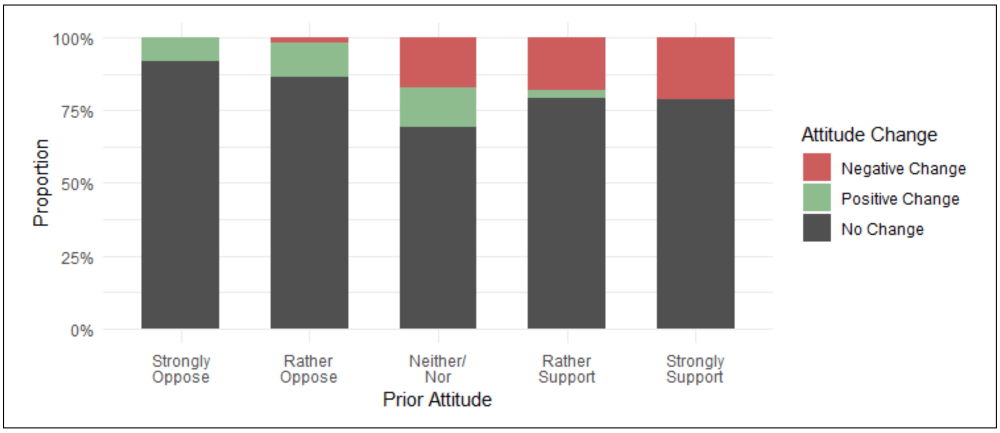
Figure 2.3: Distribution of Prior Attitude Levels by Treatment

Note. The figure displays the distribution of participants' attitudes toward AI implementation in UK courts after reading the scenario. Responses are measured on a five-point Likert scale (1 = "strongly oppose" to 5 = "strongly support") and visualized as stacked proportions. Group means are shown above each bar, and p-values are derived from MWU-tests comparing mean attitudes between treatment conditions.

2.3.2 Posterior attitudes

Overall, 476 (20.2%)[8] changed their attitudes after the discussion, with 315 shifting downward and 161 shifting upward. On average, deliberation led to a statistically significant decline in support, with mean attitudes decreasing from 3.25 to 3.14 on the 5-point scale ($p < 0.001$, paired t-test).

Figure 2.4: Distribution of attitude changes by prior attitude



Attitude changes are concentrated among participants who initially expressed more favorable views. Those who changed their attitudes showed significantly higher priors compared to

[8] This fraction is in line with results of other deliberation experiments, e.g., Hüning et al. 2022a.

non-changers (mean = 3.41 vs. 3.21; MWU-test, $p = 0.0026$). Figure 2.4 illustrates this asymmetry: participants who initially supported the tool (*Rather Support* or *Strongly Support*) were more likely to shift their attitudes after the chat (20.9%) than those who initially opposed it (*Rather Oppose* or *Strongly Oppose*) (12.5%; $\chi^2(1) = 20.25$, $p < .001$). This asymmetry is most pronounced at the extremes. Among participants who initially *strongly supported* the AI implementation, 21.4% became more skeptical after the chat. In contrast, only 8.3% of those who initially *strongly opposed* the AI shifted to a more favorable view ($\chi^2(1) = 7.79$, $p = .005$). Thus, individuals who initially expressed strong support were approximately 2.6 times more likely to change their opinion than those who initially expressed strong opposition. Figure A1 in the Appendix illustrates this graphically.

The highest rate of attitude change occurred among participants who initially held neutral views (score = 3), with 31.0% changing their opinion after the chat intervention. Among these 410 participants, 17.3% ($n = 71$) shifted toward greater skepticism, while 13.7% ($n = 56$) became more supportive. Although more participants moved in a negative direction, this difference is not statistically significant ($\chi^2(1) = 1.32$, $p = 0.25$). However, it is important to note that these are aggregated results and do not account for the influence of individual chat partners attitudes at the individual level. Since group composition was endogenous, participants were more likely to engage in discussions with others who initially held favorable attitudes.

The negative deliberation effect applies to all treatment conditions. Table 2.2 presents prior and posterior mean attitudes among participants who changed their views, along with associated p-values by treatment. While the information treatments initially led to heterogeneity in attitudes (means ranging from 3.31 to 3.52), these differences diminish after deliberation, with attitudes converging to a narrower range (2.85 to 2.89). The distribution of posterior attitudes shows significantly reduced variance compared to those of priors (Levenes F(1, 950) = 16.525, p < 0.001). This suggests that while initial information about the algorithms features and deployment, e.g., development by a public institution, increased stated support for AI, these effects diminished following deliberation, resulting in more uniform and generally more skeptical attitudes. These findings of attitudinal convergence through deliberation align with prior research showing that group discussion can moderate extreme views, foster consensus, and can increase knowledge, particularly when initial attitudes are weak or formed under uncertainty (e.g., Barabas 2004; Goeree and Yariv 2011a; Schwartzstein and Sunderam 2022; Arnesen et al. 2024).

Table 2.2: Deliberation Effect by Treatment Group

Treatment	Count	Prior Attitude	Posterior Attitude	p-Value
Restricted Input = 1	249	3.44	2.88	$p < 0.001$
Restricted Input = 0	227	3.39	2.86	$p < 0.001$
Public Oversight = 1	230	3.52	2.88	$p < 0.001$
Public Oversight = 0	246	3.31	2.85	$p < 0.001$
Judge Full Discretion = 1	246	3.47	2.85	$p < 0.001$
Judge Full Discretion = 0	230	3.35	2.89	$p < 0.001$

Note: This table reports the mean of the prior and posterior attitudes of participants who changed their attitudes after the chat discussion by treatment. P-values refer to the difference between the means of the prior and posterior attitudes and are derived from paired t-tests.

Appendix Table A11 shows corresponding logistic regression results with a binary dependent variable indicating whether a participant changed their attitude. The main insight is that those stating higher levels of AI knowledge are significantly less likely to change their attitudes following deliberation ($\beta = -0.139$, $p < 0.001$), confirming that attitude changes reflect lower attitude strength. The results also confirm that treatment assignment is not significantly linked to attitude change and participants with more favorable prior attitudes are more likely to change their attitudes ($\beta = 0.177$, $p < 0.001$). Belief in the accuracy of AI trained on restricted data also increases the likelihood of change ($\beta = 0.133$, $p < 0.001$).

2.3.3 Determinants of attitude change: Chat partners' attitudes

We now examine potential mechanisms underlying this effect, beginning with testing whether the attitudes of the randomly matched chat partners impact attitude change. The experimental design allows us to test this causally as participants were randomly assigned to chat groups within the same treatment conditions with sufficient heterogeneity in priors. Appendix Table A9 provides an overview of the group constellations and their frequencies. Most participants entered with supportive views and the likelihood of encountering supporters of AI implementation was therefore higher than encountering opponents.

Table 2.3 reports OLS estimates[9] where the dependent variable is the size of attitude change, measured as the difference between participants posterior and prior attitudes (on a 5-point Likert-scale). Columns 1 and 2 present results using the full sample. Columns 3 and 4 restrict the sample to participants with neutral prior attitudes, i.e., those that were uncertain or ambivalent when entering the chat. All models use chat partners' attitudes as independent variables. The reference group is being matched with two neutral chat partners.

[9] Ten participants dropped out after the chat, so their posterior attitudes and changes in attitudes are missing. Observations of their group members remain in the analysis to preserve the observation of that specific group constellations in the sample.

Table 2.3: Attitude of Chat Partners and Attitude Change

	DV: Attitude Change Size (Posterior - Prior)			
	(1)	(2)	(3)	(4)
One Opponent	-0.196^{**}		-0.476^{*}	
	(0.091)		(0.252)	
One Supporter	0.008		-0.358	
	(0.083)		(0.245)	
Two Opponents		-0.255^{***}		-0.742^{***}
		(0.086)		(0.246)
Two Supporters		0.079		-0.241
		(0.078)		(0.228)
Prior attitude	-0.203^{***}	-0.155^{***}		
	(0.024)	(0.019)		
Observations	745	952	131	154
Controls	YES	YES	YES	YES

Note: This table presents results from four OLS regressions with *Attitude Change* as dependent variable, which is defined as posterior attitude minus the prior attitude. The independent variables refer to the prior attitude of the two randomly matched chat partners. In all models, the reference group is encountering two neutral chat partners. *One Opponent* indicates being matched with one chat partner with a negative attitude (the other being neutral). *One Supporter* indicates being matched with one chat partner with a positive attitude (the other being neutral). *Two Opponents* (*Two Supporters*) indicates being matched with two chat partners with negative (positive) attitudes. Models (1) and (2) include all participants. Models (3) and (4) include only participants with neutral priors. All models control for participant age, gender, and education. Robust SE are in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Model 1 investigates whether being matched with one opponent or one supporter alongside one neutral affects attitude change. The results show that deliberating with one opponent significantly decreases support for algorithmic tools ($\beta = -0.196$, $p = 0.033$), while one supporter shows no significant effect ($\beta = 0.008$, $p = 0.921$). This asymmetry between meeting one opponents as compared to one supporter is statistically significant($F = 12.97$, $p =< 0.001$, linear hypothesis test). Model 2 compares participants who were matched with two opponents versus two supporters. Interacting with two opponents leads to a significantly more negative attitude change ($\beta = -0.255$, $p = 0.008$), while the effect of two supporters remains non-significant ($\beta = 0.079$, $p = 0.319$). This difference is again statistically significant ($F = 30.99$, $p < 0.001$, linear hypothesis test). Both Models 1 and 2 control for prior attitudes to account for baseline differences.

Model 3 focuses on participants who initially held neutral attitudes. We find a large but only marginally significant negative effect of being matched with one opponent on subsequent attitudes ($\beta = -0.476$, $p = 0.062$), while the effect of one supporter remains statistically insignificant ($\beta = -0.358$, $p = 0.147$). The difference between these two conditions is not statistically significant ($F = 1.24$, $p = 0.268$). However, this analysis is based on a smaller subsample, also because of the lower frequency of opponents.

In Model 4, being matched with two opponents leads to a large and negative reduction in attitudes ($\beta = -0.742$, $p = 0.005$), whereas meeting two supporters again does not show a significant effect on attitude change ($\beta = -0.241$, $p = 0.310$). This difference is statistically significant ($F = 13.86$, $p < 0.001$).

Appendix Table A10 shows the share of participants who changed their opinion after the chat, broken down by the prior attitude composition of their chat group. The likelihood of attitude change was highest when participants were matched with two opponents (26.8%) or a group including one opponent (24.9%). Appendix Table A11 shows that encountering an opponent in the chat significantly increases the odds of attitude change by 20.8% ($p = 0.092$) compared to not encountering one, while encountering a supporter significantly decreases these odds by 26.4% ($p = 0.018$). A direct comparison between these two (ignoring neutrals) shows that discussing the issue with an opponent instead of a supporter increases the odds of attitude change by 64.2%.

Taken together, the results demonstrate an asymmetry in how chat partners influence attitude formation: deliberating with opponents significantly reduces support for AI, whereas engagement with supporters shows no such an effect. Encountering even a single opponent increases skepticism, suggesting that opponents are more persuasive in the discussions. We next explore this further by applying quantitative text analysis to the chat content.

2.3.4 Chat content

We collected 15,875 individual messages from the 786 chat group conversations. The descriptive data indicates that opponents are more active in the chat. Participants with negative prior attitudes wrote more words on average (M = 99.1) than those with neutral views (M = 89.0, $p < 0.001$, MWU test) and supporters (M = 90.0, $p < 0.001$, MWU-test). The number of words written by participants with neutral and positive views does not differ significantly ($p = 0.878$). They also sent more messages on average ($M = 5.94$) compared to neutrals ($M = 5.55$, $p = 0.017$) and supporters ($M = 5.72$, $p = 0.021$). Again, there is no significant difference between the latter two groups ($p = 0.472$).

To focus the content analysis on substantive arguments and reduce noise introduced by the conversational nature of the chats, we excluded brief messages containing five words or fewer, such as greetings (hey how are you doing) or simple affirmations (I agree with your statement), from the sample.[10] The final dataset comprises 12,569 messages. The average message length is 16.7 words (median = 15; SD = 9.68).

Our text analysis classifies chat messages according to whether they express supportive or critical stances on AI implementation in the criminal justice system. To ensure robustness, we apply three distinct classification models and compare their outputs.[11] For illustrative examples of randomly sampled messages along with predicted labels and associated confidence scores, see Appendix Table A15.

The first classifier is based on BERT (Bidirectional Encoder Representations from Transformers) [12]. BERT is a transformer-based language model that captures the contextual meaning of words by jointly considering both preceding and following terms in a sentence (Devlin et al., 2019).[13] BERT generates contextualized word embeddingsnumerical representations that reflect both the meaning and context of words within a given sentence. Unlike traditional embeddings, BERTs representations vary with context. For example, the word bank is embedded differently in I went to the bank to deposit money versus The boat rested on the river bank. This capacity to disambiguate meaning is particularly valuable for our conversational dataset, where similar terms can convey different stances depending on usage. For instance, the term biased in human judges are more *biased* than ai signals support for AI, whereas ai is *biased* because it is trained on *biased* data reflects a critical view. For our analysis of conversational and unstructured chat data, we use RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019). RoBERTa follows the same architecture as BERT but uses more training data and a revised training procedure. The model has shown strong performance in handling informal language and dialogue, which fits the characteristics of our dataset (Adoma et al., 2020). We fine-tune the RoBERTa model on written responses that participants submitted prior to the chat. In these responses, participants reflected on potential benefits and drawbacks of using the AI tool. The training set helps the model learn how context-specific language signals supportive or critical views. To reduce noise, we include only responses with at least four words, excluding very short answers such as yes or I dont know. The final training set consists of 2,314 supportive responses and 2,303 critical ones. After training, we apply the model to classify each of the 12,569 chat messages as either supportive or critical of the AI tool.

[10] While expressions of agreement are relevant to our broader research question, our primary interest lies in quantifying attitude change, which we measure based on participants scale responses after the chat. Including such phrases could introduce a positivity bias in the text analysis without contributing substantive content.

[11] We do not use manual human coding for classification due to the large sample size. Coding a small subset (e.g., 500 messages) would not meaningfully validate results across the full dataset, while coding a sufficiently large portion would be prohibitively costly. Moreover, human annotation is itself subject to inconsistency and error (Bojić et al., 2025).

[12] We use Python for this task.

[13] BERT-based models are among the most widely used tools for text analysis in economics. For a review, see Ash and Hansen (2023); for recent applications, see Ferrario and Stantcheva 2022; Bursztyn et al. 2023; Moreno et al. 2025.

To further increase robustness, we train and apply two additional classifiers. The first is a Wordscore model, which is a supervised text scaling method that uses a simple bag-of-words approach (Benoit et al., 2018).[14] The model assigns a numerical score to each word based on its relative frequency in reference texts drawn from free-form responses labeled as positive or negative arguments. It then classifies new messages by averaging the scores of the words they contain, while ignoring word order and syntax. Messages with higher average scores are classified as positive; those with lower scores are classified as negative. It then classifies new messages by averaging the scores of the words they contain, while ignoring word order and syntax. Messages with higher average scores are classified as positive; those with lower scores are classified as negative. The second classifier is a Lasso-penalized logistic regression model as a weighted bag-of-words approach (Tibshirani, 1996; Ng, 2004). The model assigns a weight to each word based on how well it predicts whether a message is supportive or critical. The Lasso penalty reduces overfitting by shrinking the weights of less informative words to zero. This process selects a subset of relevant words and improves the models generalizability. Unlike Wordscore, which relies on fixed word scores derived from frequency patterns in labeled texts, the Lasso model adjusts word weights to maximize predictive accuracy. This makes Lasso more flexible but also less transparent than Wordscore. Because of the large sample size, we can restrict the text analysis to a conservative subset in which all three classifiers agree (n = 6,879). This approach reduces noise and bias in results. We also provide robustness checks based on predicted classes from each individual classifier.

We find that negative arguments appear significantly more frequently in the chat conversations, across all three classification methods ($p < 0.001$, binomial test, for each method). In the conservative subset of messages for which all three classifiers agree on the classification (n = 6,879), 66.9% of messages (4,599) are classified as negative, while 33.1% (2,271) are classified as positive. According to the BERT classification, 53.9% of all messages (6,879) are negative and 46.1% (5,899) are positive. The Wordscore model classifies 65.4% of messages as positive and 34.6% as negative. The Lasso model produces 59.9% positive and 40.1% negative classifications. This pattern is notable given that the majority of participants reported positive attitudes prior to the chat

In line with this pattern, we find that opponents contribute more arguments aligned with their stated attitudes than supporters do. Table 2.4 presents Poisson regression results where the dependent variable is the number of arguments contributed, and the key independent variables are participants prior attitudes, categorized as negative, neutral, and positive (with neutral as the reference category). The analysis is based on the conservative subset of 6,879 chat messages for which all three classification methodsBERT, Lasso, and Wordscoreagree on

[14] We use the `quanteda` package in R for this task.

the stance classification.[15] Corresponding results based on each individual classification method are reported in Appendix Tables A12, A13, and A14.

Model (1) reports results for the total number of arguments contributed, regardless of argument direction. Model (2) restricts the analysis to negative arguments, and Model (3) to positive arguments. The intercept in Model (1) is positive and significant, indicating that participants with neutral prior attitudes (the reference category) contribute actively in the chat, with an average of nearly three arguments (exp(1.060) ≈ 2.89). This shows that neutral participants do engage in the discussion and have views to express.

Models (2) and (3) break this down by argument direction. The intercepts show that neutral participants contribute more negative than positive arguments. Specifically, they contribute about 2.34 negative arguments on average (exp(0.850)) and about 1.47 positive arguments (exp(0.384)), i.e., 1.6 times more negative than positive arguments.

Table 2.4: Prior Attitudes and Number of Positive and Negative Arguments Contributed

	DV: Number of Arguments Contributed		
	(1) Total	(2) Negative Arguments	(3) Positive Arguments
Intercept	1.060^{***}	0.850^{***}	0.384^{***}
	(0.029)	(0.031)	(0.037)
Negative attitude	0.148^{***}	0.185^{***}	-0.052
	(0.036)	(0.039)	(0.049)
Positive attitude	0.036	-0.112^{***}	0.172^{***}
	(0.033)	(0.036)	(0.043)
Observations	2,221	1,941	1,394
AIC	8,479	6,521	3,920

Notes: Poisson regression models using a conservative subsample where chat message classification (positive versus negative argument) agrees across BERT, Wordscore, and LASSO methods (N=6,870 messages). The dependent variable is the number of arguments contributed during the chat and independent variables are participant's attitude prior to the chat. Measures are at the participant level, yet participants vary in their number of arguments raised and can provide arguments for both positions, positive and negative. The reference group consists of participants with a neutral prior attitude (score = 3 on a 5-point Likert scale). *Negative attitude* refers to participants who stated a 1 or 2 (opposition) before entering the chat. *Positive attitude* refers to those who stated a 4 or 5 (support). Model (1) includes all participants who contributed any argument (whose classification matched across classifiers), regardless of direction. Models (2) and (3) restrict the sample to participants who contributed at least one negative or positive argument, respectively. Participants who raised both positive and negative arguments are included in both Models (2) and (3), which explains why the sum of observations in Models (2) and (3) exceeds that of Model (1). Standard errors are clustered at the participant level and reported in parentheses. Coefficients represent log counts. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

In Model (1), the coefficient for negative prior attitudes is positive and significant ($\beta = 0.148$, $p < 0.001$), indicating that opponents contribute significantly more arguments in total than neutrals. The coefficient for positive prior attitudes is smaller and not significant ($\beta = 0.036$,

[15] These 6,879 messages were contributed by 2,221 of the original 2,358 participants, thus covering the vast majority of the sample.

$p = 0.275$), suggesting that supporters do not differ significantly from neutrals in the total number of arguments contributed.

Model (2) shows that opponents contribute significantly more negative arguments than neutrals ($\beta = 0.185$, $p < 0.001$), while supporters contribute significantly fewer ($\beta = -0.112$, $p < 0.001$). In contrast, Model (3) shows that supporters contribute significantly more positive arguments than neutrals ($\beta = 0.172$, $p < 0.001$), whereas opponents do not differ significantly from neutrals in their number of positive arguments ($\beta = -0.052$, $p = 0.288$).

Overall, these results show a strong link between participants' prior attitudes and the arguments they contribute during the chat. Moreover, opponents are more active overall, contributing a higher number of arguments. The comparison with neutral participants as reference group reveals an asymmetry: neutrals contribute more negative than positive arguments and do not differ significantly from opponents in their contribution of supportive arguments. Despite having reported a neutral stance prior to the chat, their argumentative behavior more closely resembles that of opponents than of supporters. This suggests that their neutrality may not reflect a lack of stance or disengagement. The main findings hold when using each classifier separately, although these models exhibit substantially poorer fit (Appendix Tables A12, A13, and A14). Results based on the BERT model are broadly consistent, although they indicate higher argumentative activity among supporters. All models confirm the link between prior attitudes and argument direction.

Table 2.5 reports OLS regression results examining the relationship between arguments received from chat partners and subsequent attitude change. The dependent variable is attitude change, measured as the difference between post-chat and pre-chat attitudes on a 5-point scale (ranging from 4 to +4). Model 1 includes the standardized number of positive and negative arguments received as predictors. Model 2 adds indicators for whether participants were supporters (initial score 4 or 5) or opponents (score 1 or 2) before entering the chat. Participants with neutral prior attitudes (score 3) are excluded, as we mainly focus on whether supporters and opponents differ in how they respond to arguments.

Table 2.5: Attitude Change by Arguments Received in the Chat

	(1)	(2)
Positive arguments (std.)	0.058^{***}	0.048^{*}
	(0.016)	(0.025)
Negative arguments (std.)	-0.063^{***}	-0.014
	(0.016)	(0.026)
Supporter		-0.412^{***}
		(0.110)
Supporter x positive arguments (std.)		0.014
		(0.032)
Supporter x negative arguments (std.)		-0.075^{**}
		(0.033)
Observations	1,826	1,826
R^2	0.103	0.113
Adjusted R^2	0.102	0.110

Note: OLS regression models with *attitude change* as dependent variable, measured as difference between posteriors and priors (from -4 to $+4$). *Positive/Negative arguments (std.)* are standardized (z-scored) counts of arguments received during the chat. The sample excludes participants stated stated neutral priors before the chat. *Supporter* is a dummy for participants with initially positive attitudes (score 4 or 5). Opponents serve as reference group. Sample excludes those with neutral priors. The models control for prior attitudes. Robust SE in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Model 1 estimates the baseline association between arguments received and subsequent attitude change. Receiving more positive arguments is associated with more positive attitude changes ($\beta = 0.058$, $p < 0.001$), while receiving more negative arguments are associated with negative attitude changes ($\beta = -0.063$, $p < 0.001$). The results suggest that participants updated their attitudes based on the content of the discussion, consistent with prior literature on persuasion and social learning (e.g., DellaVigna and Gentzkow 2010; Kamenica and Gentzkow 2011; Goeree and Yariv 2011b; Iaryczower et al. 2018; Schwartzstein and Sunderam 2022).

Model 2 tests heterogeneous effects based on prior attitudes (two groups: supporters vs. opponents). The reference group is opponents, for whom an increase in the number of positive arguments is marginally associated with a more favorable attitude change ($\beta = 0.048$, $p = 0.059$). This suggests that positive updates among opponents were partially driven by persuasive content. Supporters, by contrast, respond significantly more negatively when exposed to negative arguments ($\beta = -0.075$, $p = 0.023$). This suggests that supporters respond more strongly to negative arguments than opponents to positive ones. This difference in response to counterarguments is statistically significant ($F = 18.40$, $p < 0.001$, linear hypothesis test). The results are robust to ordered logistics regression models as reported in Appendix Table A23.

Topics associated with attitude change

Although not the main focus of this study, we can gain insights into the persuasive content, i.e., the topics raised in the chat that contributed to shifts in participants attitudes. To this end, we quantitatively analyze the written justifications for their positive or negative attitude change that participants provided after the discussion. We apply a text classification approach that combines term frequencyinverse document frequency (TF-IDF) representations with a logistic regression model using L1 regularization (Wang and Manning, 2012). This method identifies the most predictive words of each group, i.e. those justifying positive and those justifying negative attitude change. Term Frequency (TF) captures how often a word appears in an individual justification, while Inverse Document Frequency (IDF) reduces the weight of words that occur in many justifications. The TF-IDF score therefore emphasizes words that are both frequent in a single text and rare across others. The logistic regression model uses these weighted features and applies L1 regularization, which penalizes the inclusion of weak predictors by shrinking their coefficients to zero. As a result, the model selects a small set of words that best distinguish between justifications associated with positive and negative attitude change.

Table 2.6 lists the 30 terms most strongly associated with each group, i.e., participants who became more positive and those who became more negative, based on the absolute value of the model coefficients. The training and output include both full words and lemmatized stems. Comparing the terms across the two groups suggests that justifications associated with positive attitude change exhibit greater thematic coherence than those associated with negative change. Participants who became more supportive seem to highlight the role of AI as a tool used as support for human judges, as reflected in terms such as *alongside*, *help*, *tool*, *addition*, *aid*, *extra*,*support*, *inform* and *provide*. In other words, their increased support appears to rest on the assumption or condition that AI will complement, rather than replace, judges, and can improve human decision-making.

What is particularly striking is the magnitude of the log-likelihood coefficients in the positive group. Terms like alongside and help display substantially higher log-likelihood G^2 values than any term in the negative group. Combined with the frequent appearance of similar and connected words, this suggests that this specific theme is especially predictive and distinctive of positive attitude change.

Table 2.6: Top keywords by log-likelihood ratio in justifications of positive vs. negative attitude change

Positive Change		Negative Change	
Word stem	Log-likelihood (G^2)	Word stem	Log-likelihood (G^2)
alongside	2.999	data	0.802
help	2.666	people	0.556
benefit	1.988	negative	0.343
long	1.669	wrong	0.080
judge	1.275	develop	0.070
tool	0.904	risk	0.068
opinion	0.891	go	0.066
useful	0.864	emotion	0.064
better	0.840	bias	0.062
outweigh	0.737	individual	0.060
still	0.441	much	0.058
extra	0.341	create	0.056
positive	0.327	company	0.054
agree	0.319	someone	0.052
work	0.146	life	0.050
addition	0.128	big	0.048
aid	0.120	oversight	0.046
final	0.118	issue	0.044
inform	0.112	malfunction	0.042
never	0.109	point	0.040
support	0.103	error	0.038
good	0.099	reflect	0.036
replace	0.093	enough	0.034
assist	0.091	program	0.032
make	0.089	machine	0.030
perfect	0.087	bad	0.028
would	0.084	difficult	0.026
provide	0.082	technical	0.024
though	0.078	offend	0.022
N = 161		N = 315	

Note: This table displays the top 30 keywords predictive of justifications for positive vs. negative attitude changes after the chat, ranked by log-likelihood ratio (G^2). The values are obtained through keyness analysis comparing word frequencies in written justifications. Higher G^2 values indicate stronger statistical distinctiveness of a word for the respective group.

In contrast, words predictive of justifications for negative attitude changes show a greater topic heterogeneity. Predictive keywords include *data* and *bias*, *risk*, *malfunction*, and *error*, *oversight* and *company*, as well as *emotion*, *individual*, and *reflect*. These clusters capture distinct themes such as technical failure, fairness and data integrity, institutional accountability, and concerns related to human values and empathy. This thematic dispersion likely contributes to the lower predictive strength of individual keywords in this group. Given prior results that opponents contributed more negative arguments, this finding also suggests that opponents articulated a broader range of concerns. Many of these predictive topics were not introduced in the experimental instructions, which suggests that attitude changers learned these concerns when discussing the topic with others. These findings suggest that opponents' attitudes were shaped not only by the information provided during deliberation but also by prior knowledge, as indicated by the correlation between initial attitudes and argument stance. Their responses reflect more developed reasoning, both in the number of arguments and the diversity of topics addressed. Supporters, by contrast, appear to introduce few additional benefits beyond those presented in the experiment, for example, advantages such as improving consistency.

Appendix Tables A17 through A22 present the most frequent words used by participants, grouped by their prior attitudes (positive, negative, neutral), as well as the most frequent words they received from others in the chat, categorized by the direction of their attitude change (positive, negative, no change), respectively.

2.4 Discussion

The rapid expansion of predictive technologies such as artificial intelligence (AI) in public sector decision-making is fundamentally transforming public institutions and their procedures. Yet how citizens - those most affected by this transformation - form their attitudes toward these technologies is not well understood (De Freitas et al., 2023; Stilgoe, 2024). Even less is known about the extent to which their stated attitudes truly reflect their preferences.

To our knowledge, this is the first study to provide evidence that stated approval of AI is less robust than stated opposition. We demonstrate this using several indicators. First, effects of informational treatments that initially increase support for AI, consistent with prior survey evidence, are eliminated through peer deliberation. Second, individuals initially expressing AI support are significantly more likely to revise their stance toward skepticism. Third, individuals with an initial oppositional stance are less susceptible to counterarguments, more persuasive, and show more extensive reasoning, both in argument frequency and heterogeneity. Participants who became more skeptical referred to a broader range of topics when justifying their attitude change, suggesting opponents presented a greater diversity of considerations (Gennaioli and Shleifer, 2010).

The findings suggest that there is a substantial public uncertainty regarding the evaluation of AI adoption in critical public domains. This has important implications for the public legitimacy of this institutional transformation, suggesting that stated acceptance may be superficial and vulnerable to shifts. Such fragility could trigger future public resistance, complicate implementation efforts, and erode the sustained trust required for widespread and ethical AI integration. Moreover, the results challenge prior literature that emphasizes technical and procedural adjustments as pathways to increasing public acceptance of AI and related tools, (e.g., Dietvorst et al. 2018; Chen et al. 2025). Instead, the findings point to the need for greater investment in public information, education, and deliberative mechanisms to foster more stable and well-informed attitudes Sætra et al. 2022; Arnesen et al. 2024). At the same time, the data reveal that opposition to AI appears robust. This suggests that, in case certain AI systems offer demonstrable social benefits, realizing those gains in practice may require more sustained efforts to persuade skeptics.

The findings may have implications not only for ethical and democratic AI governance but also for other domains, including political campaigning and consumer marketing. In political campaigning, the uncertainty surrounding AI among many individuals creates a strategic opening for political actors. Unlike more established policy domains, such as climate change or migration, AI has not yet become deeply polarized and can be framed in ways that resonate across diverse constituencies. In marketing, the results align with the current wave of consumer enthusiasm surrounding AI: many consumers quickly adopt AI-branded products despite their limited understanding of the underlying technologies and their implications. The results of this study indicate that such endorsement may stem from inexperience and uncertainty, which makes it both widespread and easily swayed.

Our findings should be interpreted with caution due to several methodological limitations. As with any controlled experiment, the question of external validity remains. Although this design has the advantage of using multiple measures of attitudes, including both scaled responses and written statements, which enhances robustness, our ability to assess how these attitudes translate into real-world behavior is nonetheless limited. In particular, it remains unclear whether similar patterns would emerge in other public domains, where public attitudes toward AI may be more stable due to greater familiarity and experience with the technology. We selected the criminal justice context because it offers a realistic yet relatively unfamiliar policy domain. We assumed that most individuals have limited direct experience in this domain, which reduces the likelihood of noise and bias in their response behavior due to personal experience. Second, the sample is not representative of the broader population. Participants were recruited through Prolific, a platform known for high data quality but nonetheless subject to selection effects that undermine representativeness. Also, due to a technical issue at Prolific, our sample contains a higher proportion of female and highly educated respondents. Although we attempted to correct this imbalance through resampling, it remains present. Importantly, our analyses do not reveal significant associations between gender or education and the main outcomes, suggesting the imbalance is unlikely to drive the results. Nonetheless, given that public attitudes toward AI are likely shaped by a range of demographic and social factors, generalizability remains limited. Third, our quantitative analysis of the chat content relies on automated text classification using three distinct approaches: BERT, LASSO, and Wordscore. To ensure reliability, we restricted the analysis to the subset of messages for which all three classifiers agreed. While this conservative approach enhances internal validity, it leads to considerable data loss. Furthermore, in the absence of human-annotated ground truth, we cannot fully verify the accuracy of the remaining classifications. The opacity of some methodsparticularly BERTalso limits our ability to interpret why a given label was assigned. These challenges resemble those faced in other empirical contexts, such as noisy survey responses or inconsistent behavioral measures. Future research could address these limitations in several ways. First, deliberative studies using nationally representative samples (for a recent example in the context of AI, see Arnesen et al. (2024)) could enhance external validity and test whether the observed patterns hold across different cultural contexts and application contexts. It would also be valuable to explore how deliberative frameworks can be scaled to include broader segments of the public in discussions about AI and its governance. Online formatssuch as the one employed in this studyoffer a low-cost and scalable approach capable of reaching diverse populations across social groups and countries. Second, further exploration of AI adoption in other domains, such as education, employment, or healthcare, would help determine whether the observed patterns are specific to the criminal justice context or indicative of broader public dynamics surrounding emerging technologies. Finally, the use of large language models (LLMs) for text classification may provide a scalable and cost-effective alternative to our current ensemble method Korinek (2023); Ash et al. (2025). In future work, we plan to leverage advances in LLMs to analyze the content of deliberative discussions in greater detail, with the aim of identifying which specific arguments and discussion dynamics are most influential in shaping public attitudes toward AI.

2.5 Conclusion

Public attitudes toward AI are still forming and not yet stable. This study finds that many individuals express initially positive views, but these may reflect uncertainty rather than firm preferences. Rather than treating public support for AI as a fixed input into institutional design, our findings suggest that these attitudes are malleable and still emerging. This has implications for how scholars model public preferences, how policymakers interpret survey data, and how institutions measure legitimacy in the face of rapid technological change. Contrary to the current optimism surrounding AIs technical potential, the social foundations of its acceptance remain fragile and uneven. Understanding not only how citizens evaluate AI, but how their views evolve, is critical for ensuring that the governance of algorithmic tools remains democratically legitimate and practically effective.

2.A Appendix

Appendix A: Sessions and Sample Details

Table A1: Overview of Treatments and Fielding Dates

Treatment	Obs.	Session Dates (Year 2021)
$Judge\_full\_discretion_1, Public\_institution_0, Restricted\_inputs_0$	383	20/10, 25/10, 15/11
$Judge\_full\_discretion_1, Public\_institution_0, Restricted\_inputs_1$	363	26/10, 15/11
$Judge\_full\_discretion_1, Public\_institution_1, Restricted\_inputs_0$	353	27/10, 16/11
$Judge\_full\_discretion_1, Public\_institution_1, Restricted\_inputs_1$	354	28/10, 16/11
$Judge\_full\_discretion_0, Public\_institution_0, Restricted\_inputs_0$	353	29/10, 17/11
$Judge\_full\_discretion_0, Public\_institution_0, Restricted\_inputs_1$	350	02/11, 17/11
$Judge\_full\_discretion_0, Public\_institution_1, Restricted\_inputs_0$	358	03/11, 18/11
$Judge\_full\_discretion_0, Public\_institution_1, Restricted\_inputs_1$	350	04/11, 23/11

Note: This table presents the dates of when the treatment sessions were fielded, incl. re-sampling sessions.

Table A2: Age Distribution

Age Group (Years)	Frequencies	%
18-24	579	24.64
25-34	833	35.34
35-44	485	20.64
45-54	268	11.40
55-64	141	6.00
65-74	38	1.62
75-81	6	0.26

Note: The table shows the age distribution of participants in the final sample, with absolute frequencies and proportions across seven age categories. NAs are excluded.

Table A3: Disposable Household Income

Income Level	Count	%
less than 500	0	0
500-1000	616	27.5
1000-2000	622	27.7
2000-3000	526	23.5
3000-4000	265	11.8
4000-5000	148	6.6
more than 5000	65	2.9

Note: The table shows the household income distribution (in pounds) of of participants in the final sample (N=2,358). NAs are excluded.

Table A4: Education Distribution

Education Level	Frequency	%
Primary School	1	0
Secondary School (up to 16 years)	159	6.8
Higher Education (A-levels, etc.)	549	23.4
College or University	1,169	49.8
Post-graduate Degree	461	19.6
Prefer not to say	9	0.4

Note: This table presents the distribution of education among the participants in the final sample. NAs are excluded.

Table A5: Distribution of Attitudes before the Chat Discussion

Treatment	Str. Dis.	Disagree	Neutral	Agree	Str. Agree	Pos %	Neg %	Mean	N
Restr. inputs = 0	80	270	181	579	71	0.55	0.29	3.25	1181
Restr. inputs = 1	52	269	229	564	55	0.53	0.27	3.26	1169
Public inst. = 0	92	288	191	526	64	0.51	0.32	3.16	1161
Public inst. = 1	40	251	219	617	62	0.57	0.24	3.34	1189
Judge full d. = 0	64	243	208	578	64	0.55	0.26	3.29	1157
Judge full d. = 1	68	296	202	565	62	0.52	0.30	3.22	1193

Note: This table presents the distribution of attitudes in each treatment condition before the chat (as counts). *Pos %* represents the proportion of participants with positive attitudes (*agree* or *strongly agree*), while *Neg %* represents the proportion of participants with negative attitudes (*disagree* or *strongly disagree*). The question participants responded to was: *How strongly do you agree with this statement: The AI, as explained in this scenario, should be used in criminal courts?*

Table A6: Distribution of Prior Attitudes and Direction of Attitude Changes

Prior Attitude	Count (Percentage)	Positive Change	Negative Change
1	133 (5.60%)	11 (8.27%)	0 (0.00%)
2	541 (22.90%)	63 (11.65%)	10 (1.85%)
3	410 (17.40%)	56 (13.66%)	71 (17.32%)
4	1148 (48.70%)	31 (2.70%)	207 (18.03%)
5	126 (5.30%)	0 (0.00%)	27 (21.43%)

Note: This table reports the distribution of participants' prior attitudes (measured on a 15 scale, where 1 = strongest opposition and 5 = strongest support) and their subsequent attitude changes after deliberation. *Count (Percentage)* refers to the number and share of participants at each attitude level. *Positive Change* and *Negative Change* indicate the number and percentage of participants whose attitudes shifted upward or downward after the chat, respectively.

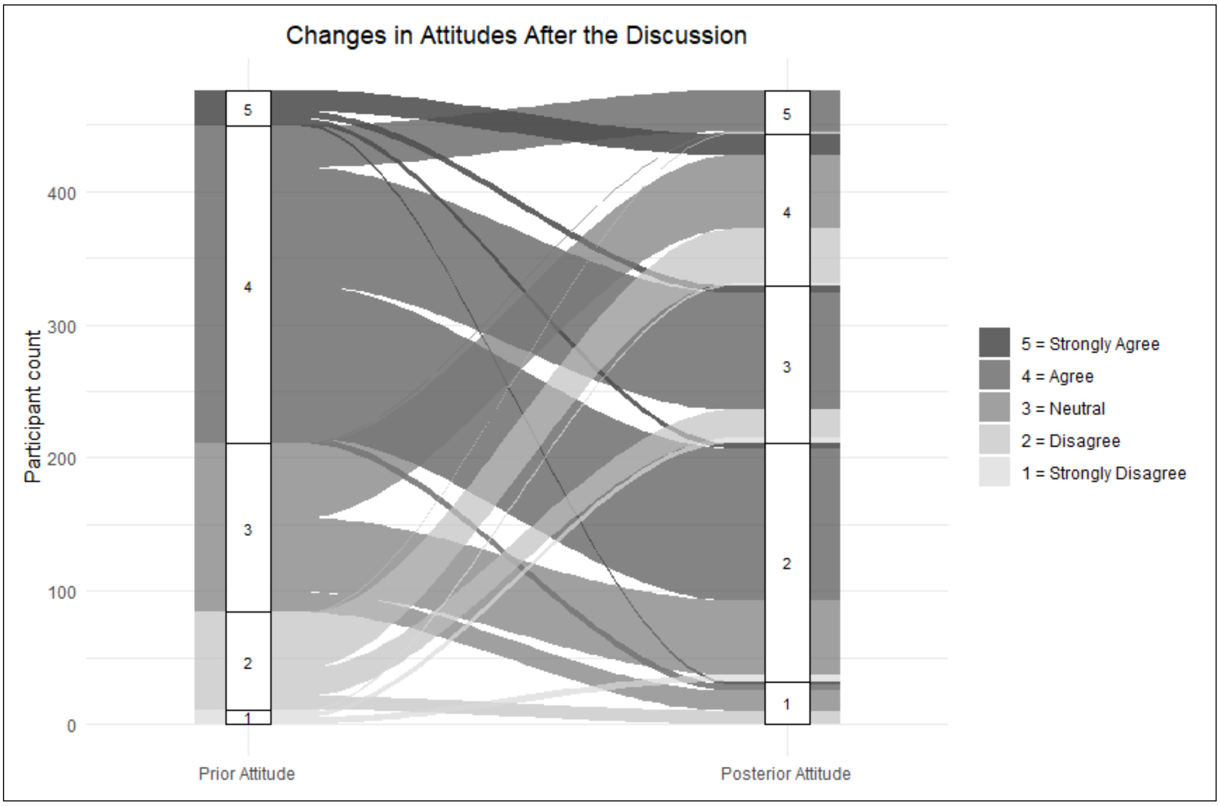


Figure A1: Alluvial Plot of Attitude Changes

Table A9: Chat Group Constellations by Prior Attitudes

Chat Group Constellation	Frequency	Relative Frequency (%)
negative positive positive	197	25.06
negative neutral positive	134	17.05
positive positive positive	128	16.28
neutral positive positive	110	13.99
negative negative positive	97	12.34
neutral neutral positive	45	5.73
negative negative neutral	35	4.45
negative negative negative	21	2.67
negative neutral neutral	16	2.04
neutral neutral neutral	3	0.38
Total	786	100.00

Note: This table reports the absolute and relative frequency of each chat group constellation based on the attitudes of participants before entering the chat conversations (positive, neutral, negative).

Table A7: Attitudes Before Chat

	DV: Prior Attitudes	
	(1)	(2)
Judge full discretion	0.186**	0.204***
	(0.078)	(0.079)
Public institution	0.338***	0.364***
	(0.078)	(0.079)
Restricted inputs	−0.001	0.003
	(0.078)	(0.078)
AI knowledge	0.086	0.071
	(0.055)	(0.056)
AI knowledge (confidence)	−0.067	−0.074
	(0.058)	(0.059)
Belief: Judge bias	0.143**	0.103*
	(0.056)	(0.058)
Belief: AI less biased than judge	0.571***	0.592***
	(0.042)	(0.042)
Belief: Restricted AI is accurate	0.382***	0.374***
	(0.038)	(0.038)
Age		−0.015***
		(0.003)
Female		−0.034
		(0.090)
Education		0.042
		(0.048)
Observations	2,358	2,348
AIC	5906.60	5866.09

Notes: Ordered logistic regression predicting stated support for the implementation of AI in court decisions, measured on a 5-point Likert scale (1 = strongly oppose, 5 = strongly support). *Judge full discretion*, *Public institution*, and *Restricted inputs* are treatment indicators (1 = feature present in scenario). *AI knowledge* and *AI knowledge (confidence)* are self-assessed measures of understanding of AI (absolute and relative). *Judge bias*, *AI less biased than judge*, and *Restricted AI is accurate* are belief measures. Model 2 includes demographic controls. *Education* is an ordered categorical variable. Standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A8: Attitude Change (Binary)

	DV: Attitude Change (Binary)		
	(1)	(2)	(3)
Public institution	-0.160	-0.169	-0.152
	(0.104)	(0.104)	(0.105)
Judge full discretion	0.112	0.100	0.113
	(0.103)	(0.104)	(0.104)
Restricted inputs	0.124	0.129	0.130
	(0.104)	(0.104)	(0.104)
AI knowledge	-0.133**	-0.140**	-0.136*
	(0.060)	(0.061)	(0.064)
AI knowledge relative to others	0.043	0.042	0.040
	(0.076)	(0.076)	(0.076)
AI attitude general (positive)	-0.002	-0.001	0.003
	(0.084)	(0.084)	(0.084)
AI attitude general (negative)	-0.119	-0.121	-0.133
	(0.088)	(0.088)	(0.088)
Belief: Judge bias		0.029	-0.017
		(0.072)	(0.074)
Belief: AI less biased than judge		-0.103*	-0.083
		(0.054)	(0.054)
Belief: Accuracy AI with restricted input		0.148***	0.136***
		(0.051)	(0.051)
Prior attitudes	0.219***	0.217***	0.201***
	(0.054)	(0.057)	(0.057)
Age			-0.019***
			(0.005)
Female			0.118
			(0.123)
Education			0.005
			(0.064)
Observations	2350	2350	2348
AIC	2358.0	2350.3	2337.7

Notes: The dependent variable is *Attitude Change (Binary)*, coded as 1 if a participant changed their attitude after the chat. Model 1 includes treatment indicators, self-assessed AI knowledge (actual and relative), general attitudes toward AI, and prior attitudes. Model 2 adds beliefs about judges and AI. Model 3 includes demographics. *Belief: Judge bias* captures perceptions that human judges are influenced by prejudice. *Belief: AI less biased than judge* reflects whether participants believe AI is fairer than judges. *Belief: Accuracy AI with restricted input* measures the belief that AI remains accurate even when sensitive attributes are excluded from its inputs. Robust standard errors are in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Attitude Change by Chat Composition

Chat Composition	N	Changed	Change Rate	Difference (pp)
negative negative	194	52	0.268	2.52
negative neutral	233	58	0.249	0.61
neutral neutral	70	17	0.243	0.00
negative positive	722	150	0.208	−3.51
neutral positive	442	86	0.195	−4.83
positive positive	689	113	0.164	−7.89
Total	2,350	476	0.203	—

Note: This table reports the number and share of participants who changed their opinion after the chat, by chat partner composition. Change is defined as a difference between the pre- and post-chat attitude (measured on a 5-point Likert scale). The column "Difference (pp)" indicates the change in percentage points compared to the baseline case of being matched with two neutrals.

Table A11: Determinants of Attitude Change (Binary)

	Attitude Change (Binary)
Intercept	-1.874^{***}
	(0.224)
Encountered Opponent	0.189^{*}
	(0.112)
Encountered Supporter	-0.307^{**}
	(0.129)
Prior Attitude	0.194^{***}
	(0.051)
Observations	2342
AIC	2348.4

Note: Logistic regression model with *Attitude Change* as dependent variable, measured as a binary indicator (1 for change, 0 for no change). *Encountered Opponent* is a binary variable indicating if the participant was paired with at least one opponent in the chat. *Encountered Supporter* is a binary variable indicating if the participant was paired with at least one supporter. The reference group for these variables is encountering a neutral participant. Standard errors are reported in parentheses. Coefficients represent log-odds. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A12: Prior Attitude and Number of Arguments Contributed (BERT)

	Dependent variable: Number of Arguments		
	(1) Total	(2) Negative	(3) Positive
Intercept	1.630***	1.075***	1.025***
	(0.023)	(0.028)	(0.029)
Negative attitude	0.092***	0.182***	0.026
	(0.029)	(0.035)	(0.035)
Positive attitude	0.181***	0.015	0.219***
	(0.024)	(0.029)	(0.030)
Observations	2339	2151	2123
AIC	10830.23	8249.78	7678.92

Note: Poisson regression models using based on BERT classifier. The dependent variable is the number of arguments contributed during the chat. The reference group consists of participants with a neutral prior attitude (score = 3 on a 5-point Likert scale). *Negative attitude* refers to participants who stated a 1 or 2 (opposition) before entering the chat. *Positive attitude* refers to those who stated a 4 or 5 (support). Model (1) includes all participants who contributed any argument, regardless of direction. Models (2) and (3) restrict the sample to participants who contributed at least one negative or positive argument, respectively. Participants who raised both positive and negative arguments are included in both Models (2) and (3), which explains why the sum of observations in Models (2) and (3) exceeds that of Model (1). Standard errors are clustered at the participant level and reported in parentheses. Coefficients represent log counts. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A13: Prior Attitude and Number of Arguments Contributed (Lasso)

	Dependent variable: Number of Arguments		
	(1) Total	(2) Negative	(3) Positive
Intercept	1.630***	1.220***	0.783***
	(0.023)	(0.028)	(0.030)
Negative attitude	0.092***	0.154***	-0.037
	(0.029)	(0.035)	(0.039)
Positive attitude	0.043	-0.095***	0.199***
	(0.027)	(0.033)	(0.035)
Observations	2339	2215	2063
AIC	10830.23	8842.83	7046.11

Note: Poisson regression models based on Lasso (glmnet) classifier. The dependent variable is the number of arguments contributed during the chat. The reference group consists of participants with a neutral prior attitude (score = 3 on a 5-point Likert scale). *Negative attitude* refers to participants who stated a 1 or 2 (opposition) before entering the chat. *Positive attitude* refers to those who stated a 4 or 5 (support). Model (1) includes all participants who contributed any argument, regardless of direction. Models (2) and (3) restrict the sample to participants who contributed at least one negative or positive argument, respectively. Participants who raised both positive and negative arguments are included in both Models (2) and (3), which explains why the sum of observations in Models (2) and (3) exceeds that of Model (1). Standard errors are clustered at the participant level and reported in parentheses. Coefficients represent log counts. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A14: Prior Attitude and Number of Arguments Contributed (Wordscore)

	Dependent variable: Number of Arguments		
	(1) Total	(2) Negative	(3) Positive
Intercept	1.630***	1.290***	0.708***
	(0.023)	(0.027)	(0.029)
Negative attitude	0.092***	0.142***	-0.028
	(0.029)	(0.033)	(0.040)
Positive attitude	0.043	-0.065*	0.162***
	(0.027)	(0.031)	(0.034)
Observations	2339	2242	1956
AIC	10830.23	9118.62	6390.77

Note: Poisson regression models based on Wordscore classifier. The dependent variable is the number of arguments contributed during the chat. The reference group consists of participants with a neutral prior attitude (score = 3 on a 5-point Likert scale). *Negative attitude* refers to participants who stated a 1 or 2 (opposition) before entering the chat. *Positive attitude* refers to those who stated a 4 or 5 (support). Model (1) includes all participants who contributed any argument, regardless of direction. Models (2) and (3) restrict the sample to participants who contributed at least one negative or positive argument, respectively. Participants who raised both positive and negative arguments are included in both Models (2) and (3), which explains why the sum of observations in Models (2) and (3) exceeds that of Model (1). Standard errors are clustered at the participant level and reported in parentheses. Coefficients represent log counts. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A15: Examples of Chat Messages and Corresponding Classifier Predictions

Message	**Wordscore** Prediction	**Wordscore** Conf. Score	**BERT** Prediction	**BERT** Conf. Score	**Lasso** Prediction	**Lasso** Conf. Score
judges are highly intelligent and well trained and must undergo years of training and sit difficult exams	negative	-0.074	negative	0.54	negative	0.57
you are not waiting for me to finish responding to that question	positive	0.290	negative	0.54	negative	0.56
my concern is that private companies might not be accountable	negative	-0.439	negative	0.81	negative	0.78
its impossible for there not to be bias as judges are only human	positive	0.057	negative	0.74	positive	0.73
ai can be used to support judges not replace them	positive	0.120	positive	0.89	positive	0.76
i dont think we should let ai take the lead	negative	-0.182	negative	0.91	negative	0.74
resources could be better allocated with ai	positive	0.349	positive	0.77	positive	0.66
theres still value in human judgme in court decisions	positive	0.311	negative	0.86	positive	0.75
ai can reinforce historical biases in the justice system	negative	-0.461	negative	0.90	negative	0.77
humans can also be inconsistent or emotional in judgments	positive	0.298	negative	0.67	positive	0.74
i mean a lot of processes are being automised which does help as well	positive	0.278	positive	0.88	positive	0.75
i believe it will help judges	positive	0.726	positive	0.94	positive	0.82
we need to be cautious about how much we rely on ai	negative	-0.422	negative	0.84	negative	0.70
algorithms can't understand context like a person can	negative	-0.391	negative	0.86	negative	0.77
ai in sentencing might lead to more consistency	positive	0.471	positive	0.88	positive	0.76

Note: Randomly sampled chat messages with predicted class labels and confidence scores from three classification models.

Table B9 (continued): Examples of Chat Messages and Corresponding Classifier Predictions

Message	**Wordscore** Prediction	**Wordscore** Conf. Score	**BERT** Prediction	**BERT** Conf. Score	**Lasso** Prediction	**Lasso** Conf. Score
yeah the i think judges should just be vetted a bit more	positive	0.106	negative	0.50	positive	0.56
it would help reduce court waiting times	positive	0.769	positive	0.89	positive	0.96
people could also be very against technology as they dont trust it	negative	-0.484	negative	0.76	negative	0.95
yes facial movement detector is great	negative	-0.118	positive	0.92	none	0.51
i think the positives would outweigh the negatives	positive	0.069	positive	0.59	positive	0.51
how is being released early a good thing	negative	-0.073	positive	0.87	none	0.53
yeah its interesting i find ai kinda scary though	negative	-0.331	negative	0.79	negative	0.97
who will oversee all of this	negative	-1.000	negative	0.58	negative	0.64
definitely and theres little information on who this private company is	negative	-0.552	negative	0.58	negative	0.93
enjoy the rest of your day	negative	-0.115	positive	0.62	positive	0.68
i think it would have mostly a negative effect	negative	-0.332	negative	0.76	negative	0.96
should prison behavior data be included in the ai data set as well to counter that	negative	-0.275	negative	0.81	negative	0.90
sorry it s late please excuse my poor grammar	negative	-0.323	negative	0.60	negative	0.68
i think ai will be more successful at determining whether a criminal will reoffend	negative	-0.047	positive	0.81	negative	0.66
i think judges should take more training in being biased	negative	-0.135	negative	0.60	positive	0.66

Note: Randomly sampled chat messages with predicted class labels and confidence scores from three classification models.

Table A17: Top 50 Words in Messages Sent by Participants with Positive Prior Attitudes

Rank	**Word**	**Absolute Frequency**	**Relative Frequency**
1	ai	1674	0.047279
2	judge	940	0.026548
3	positive	846	0.023894
4	decision	740	0.020900
5	agree	598	0.016889
6	bias	594	0.016776
7	data	568	0.016042
8	human	499	0.014093
9	judges	494	0.013952
10	negative	343	0.009687
11	people	334	0.009433
12	decisions	325	0.009179
13	time	319	0.009010
14	system	311	0.008784
15	biased	249	0.007033
16	person	246	0.006948
17	impact	235	0.006637
18	yeah	234	0.006609
19	based	219	0.006185
20	feel	207	0.005846
21	information	199	0.005620
22	justice	193	0.005451
23	true	183	0.005168
24	its	176	0.004971
25	final	171	0.004830
26	money	164	0.004632
27	tool	147	0.004152
28	don	146	0.004123
29	biases	145	0.004095
30	lot	140	0.003954
31	factors	127	0.003587
32	implications	127	0.003587
33	crime	126	0.003559
34	account	123	0.003474
35	court	118	0.003333
36	negatives	117	0.003304
37	save	116	0.003276
38	idea	113	0.003191
39	release	112	0.003163
40	effect	105	0.002966
41	fair	102	0.002881
42	prisoner	100	0.002824
43	process	100	0.002824
44	race	100	0.002824
45	reoffending	100	0.002824
46	humans	99	0.002796
47	reduce	99	0.002796
48	past	95	0.002683
49	prison	94	0.002655
50	individual	93	0.002627

Note: Top 50 words sent by participants with positive prior attitudes. Stopwords are removed.

Table A18: Top 50 Words in Messages Sent by Participants with Negative Prior Attitudes

Rank	**Word**	**Absolute Frequency**	**Relative Frequency**
1	ai	1092	0.050904
2	judge	388	0.018087
3	data	379	0.017667
4	bias	344	0.016036
5	decision	331	0.015430
6	negative	326	0.015197
7	human	313	0.014591
8	agree	309	0.014404
9	judges	268	0.012493
10	people	259	0.012073
11	system	234	0.010908
12	decisions	221	0.010302
13	positive	207	0.009649
14	biased	177	0.008251
15	time	165	0.007692
16	person	146	0.006806
17	justice	145	0.006759
18	don	135	0.006293
19	impact	127	0.005920
20	based	126	0.005874
21	biases	123	0.005734
22	money	113	0.005268
23	feel	112	0.005221
24	company	103	0.004801
25	information	101	0.004708
26	yeah	90	0.004195
27	individual	89	0.004149
28	account	86	0.004009
29	humans	86	0.004009
30	idea	83	0.003869
31	implications	82	0.003822
32	release	79	0.003683
33	private	77	0.003589
34	true	76	0.003543
35	previous	73	0.003403
36	its	72	0.003356
37	cost	69	0.003216
38	computer	68	0.003170
39	risk	68	0.003170
40	factors	66	0.003077
41	past	63	0.002937
42	process	62	0.002890
43	public	61	0.002844
44	save	60	0.002797
45	prisoner	59	0.002750
46	crime	57	0.002657
47	historical	57	0.002657
48	prison	56	0.002610
49	fair	55	0.002564
50	lot	55	0.002564

Note: Top 50 words used by participants with Negative Prior Attitudes. Stopwords are removed.

Table A19: Top 50 Words in Messages Sent by Participants with Neutral Prior Attitudes

Rank	**Word**	**Absolute Frequency**	**Relative Frequency**
1	ai	618	0.054411
2	judge	221	0.019458
3	positive	198	0.017433
4	decision	192	0.016904
5	agree	191	0.016816
6	human	185	0.016288
7	bias	175	0.015408
8	data	172	0.015144
9	negative	147	0.012942
10	judges	134	0.011798
11	people	110	0.009685
12	decisions	108	0.009509
13	time	107	0.009421
14	system	86	0.007572
15	biased	75	0.006603
16	person	75	0.006603
17	feel	72	0.006339
18	information	68	0.005987
19	based	66	0.005811
20	yeah	64	0.005635
21	don	55	0.004842
22	individual	55	0.004842
23	justice	52	0.004578
24	money	52	0.004578
25	biases	47	0.004138
26	court	47	0.004138
27	true	47	0.004138
28	account	43	0.003786
29	final	43	0.003786
30	its	42	0.003698
31	process	42	0.003698
32	prisoner	41	0.003610
33	crime	40	0.003522
34	humans	40	0.003522
35	implications	40	0.003522
36	factors	39	0.003434
37	idea	38	0.003346
38	impact	38	0.003346
39	lot	37	0.003258
40	tool	35	0.003082
41	bad	34	0.002993
42	company	33	0.002905
43	release	33	0.002905
44	computer	31	0.002729
45	dont	31	0.002729
46	previous	31	0.002729
47	private	30	0.002641
48	reduce	30	0.002641
49	save	30	0.002641
50	cost	29	0.002553

Note: Top 50 words sent by participants with neutral prior attitudes. Stopwords are removed.

Table A20: Top 50 Words in Messages Received by Positive Attitude Changers (N = 161)

Rank	**Word**	**Absolute Frequency**	**Relative Frequency**
1	think	521	0.037474
2	ai	489	0.035172
3	judge	239	0.017191
4	agree	207	0.014889
5	positive	202	0.014529
6	yes	181	0.013019
7	decision	165	0.011868
8	human	152	0.010933
9	data	143	0.010286
10	bias	141	0.010142
11	good	135	0.009710
12	also	118	0.008487
13	judges	112	0.008056
14	used	105	0.007552
15	negative	100	0.007193
16	people	98	0.007049
17	make	87	0.006258
18	time	86	0.006186
19	system	86	0.006186
20	decisions	82	0.005898
21	need	77	0.005538
22	cases	76	0.005466
23	true	76	0.005466
24	see	73	0.005251
25	biased	68	0.004891
26	like	65	0.004675
27	yeah	65	0.004675
28	feel	64	0.004603
29	person	64	0.004603
30	less	64	0.004603
31	justice	63	0.004531
32	definitely	57	0.004100
33	still	56	0.004028
34	help	55	0.003956
35	impact	52	0.003740
36	use	52	0.003740
37	say	50	0.003596
38	information	48	0.003452
39	one	48	0.003452
40	point	46	0.003309
41	though	46	0.003309
42	making	45	0.003237
43	based	45	0.003237
44	hi	44	0.003165
45	way	44	0.003165
46	final	43	0.003093
47	take	43	0.003093
48	case	42	0.003021
49	maybe	42	0.003021
50	implications	41	0.002949

Note: Top 50 words received by participants with positive attitude change. Stopwords are removed.

Table A21: Top 50 Words in Messages Received by Negative Attitude Changers

Rank	Word	Absolute Frequency	Relative Frequency
1	ai	1026	0.034886
2	think	950	0.032302
3	judge	412	0.014009
4	agree	399	0.013567
5	decision	347	0.011799
6	positive	329	0.011187
7	yes	319	0.010847
8	bias	315	0.010711
9	data	307	0.010440
10	human	298	0.010133
11	negative	271	0.009215
12	judges	253	0.008603
13	also	248	0.008433
14	people	241	0.008194
15	like	213	0.007242
16	good	212	0.007208
17	system	203	0.006902
18	decisions	189	0.006426
19	make	189	0.006426
20	used	186	0.006324
21	time	165	0.005610
22	cases	163	0.005542
23	true	150	0.005100
24	case	144	0.004896
25	biased	143	0.004862
26	need	132	0.004488
27	yeah	131	0.004454
28	take	128	0.004352
29	see	126	0.004284
30	person	123	0.004182
31	use	121	0.004114
32	feel	120	0.004080
33	justice	116	0.003944
34	making	115	0.003910
35	still	112	0.003808
36	based	109	0.003706
37	impact	106	0.003604
38	help	104	0.003536
39	exactly	101	0.003434
40	someone	98	0.003332
41	money	96	0.003264
42	one	96	0.003264
43	maybe	95	0.003230
44	much	94	0.003196
45	etc	93	0.003162
46	using	90	0.003060
47	way	89	0.003026
48	definitely	88	0.002992
49	point	87	0.002958
50	really	86	0.002924

Note: Top 50 words received by participants with negative attitude change. Stopwords are removed.

Table A22: Top 50 Words in Messages Received by Participants without Change

Rank	Word	Absolute Frequency	Relative Frequency
1	think	5552	0.034728
2	ai	5289	0.033083
3	judge	2471	0.015456
4	agree	2216	0.013861
5	positive	2100	0.013135
6	decision	2020	0.012635
7	yes	1921	0.012016
8	data	1802	0.011271
9	bias	1780	0.011134
10	human	1546	0.009670
11	judges	1435	0.008976
12	also	1368	0.008557
13	negative	1339	0.008375
14	good	1193	0.007462
15	make	1140	0.007131
16	people	1091	0.006824
17	decisions	1043	0.006524
18	like	1008	0.006305
19	system	975	0.006099
20	time	951	0.005948
21	used	893	0.005586
22	see	879	0.005498
23	biased	801	0.005010
24	cases	781	0.004885
25	yeah	778	0.004866
26	person	751	0.004697
27	true	724	0.004529
28	need	723	0.004522
29	making	692	0.004328
30	based	670	0.004191
31	impact	646	0.004041
32	case	642	0.004016
33	help	619	0.003872
34	information	609	0.003809
35	still	606	0.003791
36	justice	601	0.003759
37	feel	600	0.003753
38	take	595	0.003722
39	use	581	0.003634
40	definitely	570	0.003565
41	one	538	0.003365
42	biases	535	0.003346
43	money	535	0.003346
44	point	527	0.003296
45	maybe	521	0.003259
46	way	501	0.003134
47	less	500	0.003127
48	sure	463	0.002896
49	etc	462	0.002890
50	say	457	0.002859

Note: Top 50 words received by participants without attitude change. Stopwords are removed.

Table A23: Attitude Change by Arguments Received (Ordered Logistic Regression)

	(1)	(2)
Positive arguments (std.)	0.263^{***}	0.222^{**}
	(0.063)	(0.101)
Negative arguments (std.)	-0.256^{***}	-0.144
	(0.061)	(0.117)
Supporter		-1.987^{***}
		(0.453)
Supporter Œ positive arguments (std.)		0.063
		(0.129)
Supporter Œ negative arguments (std.)		-0.152
		(0.136)
Prior attitude	-0.790^{***}	-0.073
	(0.066)	(0.179)
Observations	1,826	1,826
AIC	2,492	2,476

Note: Ordered logistic regression models predicting *attitude change* (ordinal outcome: -4 to $+4$). *Positive/Negative arguments (std.)* are standardized (z-scored) counts of arguments received during the chat. The sample excludes participants stated stated neutral priors before the chat. *Supporter* is a dummy for participants with initially positive attitudes (score 4 or 5). Opponents serve as reference group. Sample excludes those with neutral priors. The models control for prior attitudes. Robust SE in parentheses. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.A.1 Experimental Instructions

Slide 1

Welcome to this survey. Thank you for your participation!

This survey is conducted by researchers at the University of Hamburg.

We are interested in your opinion on the application of artificial intelligence (AI) in court decisions. You do not need any pre-knowledge on AI or law.

Later, you will be asked to chat with other participants about the scenario presented to you. Including the chat discussion, the study will last approx. 30 minutes. Payment will be paid out only for completed surveys and chat participation. Please note that during the study, you cannot go back to previous questions.

Your answers will be kept confidential and anonymous. You can withdraw from the survey at any time. Please note that only completed surveys are valuable for research.

Your answers can make an impact: The UK government may consult the aggregated results of this study in their decision-making process on the implementation of AI systems in criminal courts.

- Yes, I'd like to participate.
- No, I don't want to participate.

Slide 2

Please answer the following questions: How familiar are you with the topic of Artificial Intelligence (AI)?

Likert-scale format: 1=very familiar to 5=not familiar at all

- How familiar are you with the topic of Artificial Intelligence (AI)?
- How familiar do you think other UK citizens are with the AI topic?

Slide 3

Please answer the following questions:

Likert-scale format: 1=very strong to 5=not strong at all

- How strong is your positive attitude toward AI?
- How strong is your negative attitude toward AI?

Slide 5

Artificial intelligence (AI) refers to computer systems that learn from data to make predictions about uncertain events. In decision-making, AI uses large amounts of past data to predict how someone might behave in the future.

It compares an individuals characteristics to patterns found in others and uses that information to generate a prediction, which can then support human decision-making.

Slide 6

Imagine a scenario in the criminal justice system: the AI system supports judges in making decisions about early release for prisoners. The AI system analyzes prisoners' characteristics, such as age and criminal record, and compares them to thousands of past cases. Based on these comparisons, the AI system predicts the likelihood that the prisoner will commit another crime if released early. The judge then receives a probability of the prisoner's reoffending behavior. The judge always makes the final decision.

Slide 7

How is Artificial Intelligence (AI) used in this scenario?(only one answer is correct)

- Revealing criminals identity.
- Predicting the probability that a prisoner reoffends.
- Communicating between prisoner and lawyer.
- Supporting in the investigation if a defendant is guilty of a crime.

Slide 8

Random factors and personal biases can influence how judges make early release decisions.

Research shows that factors such as the weather, a judges mood, or characteristics of the prisonersuch as gender, skin color, or religioncan affect a judges decision, even if they are not legally relevant. These influences may lead to unfair or inconsistent outcomes.

Sources: Englich Soder (2009), Danzinger et al. (2011), and Davids (2017).

Slide 9

Please answer the following questions: Judges are required to base their decisions on evidence relevant to the specific case at hand. (Being "biased" means being influenced by random factors or prejudices when judging a case.)

Likert-scale: 1=always biased to 5=never biased

Slide 19

Biased judgment of judges in the past may become part of the data that AI uses for its predictions. The AI can learn these biases.

Slide 20

AI makes predictions based on historic data. This data can carry prejudices and discriminatory decisions made by judges. Do you think AI using all past data makes more or less biased predictions than the average judge?

Likert-scale: 1=a lot more biased to 5=a lot less biased

Slide 21

Alternative Method: Restrict AI to exclude data related to common prejudices.
This means excluding attributes like race, gender, and religion, from the data.
However, this could lead to reduced accuracy of the predictions. The predictions may be less informative about the prisoner.

Slide 22

AI can be restricted to exclude data related to the most common prejudices. Do you think AI that is based on such a restricted database generates more or less accurate predictions, compared to AI that is based on an unrestricted database i.e. use of all possible information about the prisoner?

Likert-scale: 1=a lot more accurate to 5=a lot less accurate

Slide 23

You are now presented a scenario, which you are asked to evaluate later on.
There is no correct or wrong answer. We are interested in your subjective assessment.
Results from studies like this one often inform real-world policy-making.
Please read all information carefully.

Slide 24

Several industrial countries worldwide are increasingly using artificial intelligence (AI) to assist criminal courts in making early release decisions. These AI systems can predict the likelihood that prisoners will reoffend if released from prison. Each prisoner is assigned a low, medium, or high risk of reoffending score. The AI makes predictions based on the prisoner's personal characteristics (e.g., criminal history, family relationships, social exclusion, and attitudes) using sophisticated machine learning models linked to a national database on criminal behavior. Predicting future behavior, including a prisoners probability of reoffending, is generally one of the main factors judges consider when making early release decisions.

- The judge must incorporate the reoffending risk score in the final decision. [Treatment variation: The judge has free discretion in using or ignoring the risk score.]
- The AI is developed by a private company. As the tool is proprietary to the company and protected by trade secrecy laws, it is subject to very limited federal and public

oversight. Nobody can access the inner functioning of the tool, including how it weighs certain characteristics, and about the data it uses for prediction.[Treatment variation: The AI tool is developed by a public institution. Independent experts from these public institutions can access the system and review how it works, including how it weighs different characteristics and what data it uses to make predictions.]

- The AI is built upon a vast amount of historic and current legal data. Therefore, it can carry in itself potential judgment biases (e.g. with regard to gender, religion) and inequalities in the society, thereby potentially amplifying discrimination in court judgments. Although it is technically possible to counter such biases underlying the data, any such intervention is avoided as this could require dropping valuable information. [Treatment variation: The AI is built upon a vast amount of historic and current legal data. Therefore, it can carry in itself potential judgment biases (e.g. with regard to gender, religion) and inequalities in the society, thereby potentially amplifying discrimination in court judgments. The AI excludes such variables from its data. However, such interventions could lead to less informative predictions.]

Slide 25

Your opinion
How strongly do you agree with the following statement?: "The AI, as explained in the scenario, should be implemented in criminal courts."

Likert-scale: 1=strongly agrree to 5=strongly disagree

Slide 26

Your opinion
Would this AI, as described in the scenario, have positive effects? If yes, why? *Free-form text box here*
Would this AI, as described in the scenario, have negative effects? If yes, why? *Free-form text box here*

Slide 27

Chat discussion Next, you are randomly matched with two other participants of the survey to discuss in a chat the whether this AI should be implemented in UK courts or not. The chat will last 7 minutes. Chat rules:

- Do NOT reveal your identity.
- Do NOT insult, harass, or isolate other participants in any way.
- Do NOT lie about the procedures of the survey.

Slide 28

Waiting Room

A chat group is now being created. You have a max. waiting time of 7 minutes. Please stay in front of the screen. The chat will start as soon as enough participants have entered the chat section. Please watch this browser window during the waiting period. If your device and browser have a speaker or audio setup enabled, you can hear a sound as soon as the chat with your assigned partners is ready. Please wait. The chat will start soon.

Slide 29

Chat discussion

Slide 30

Would you like to change your previous answer?

How strongly do you agree with the following statement?: "The AI, as explained in the scenario, should be implemented in criminal courts." *Previous answer is shown here.*

- Yes, I would like to change my answer
  - (Likert scale: 1=strongly agree to 5=strongly disagree)
- No, I don't want to change my answer.

(If "Yes", a free-form text box appears asking: Why did you change your answer?)

Slide 31

With how many chat partners did you agree?

With how many chat partners did you disagree?

Do you think you convinced any chat partner of your viewpoint?

For each of these questions the following scale is used:

- 0 (none)
- 1 (with one chat partner)
- 2 (with both chat partners)

Slide 32

Demographic questions

Slide 33

This is the end of the survey. Thank you very much for your participation. [*Redirect to Prolific*]

Chapter 3.

Algorithmic Fairness and Human Discrimination

Abstract

Fairness constraints in algorithm design aim to reduce discrimination. Their impact, however, also depends on the adoption of the algorithm by human-decision makers as they typically retain full authority in high-stakes contexts. In a hiring experiment, I first find suggestive evidence that protecting group membership in algorithmic predictions leads individuals to be more conservative in updating their beliefs about candidates based on these predictions. I then find a significant increase in discrimination in their hiring of candidates under this algorithm, driven by those who initially believe that group membership predicts performance. Finally, irrespective of the algorithm features, about 26% of participants make hiring decisions that cannot be explained by beliefs and are likely based on taste. These results suggest that algorithmic fairness features can paradoxically exacerbate human discrimination based on statistical beliefs by hindering adoption and, unsurprisingly, remain orthogonal to taste-based discrimination.

Keywords: Algorithmic Fairness, Belief Updating, Hiring Experiment

JEL Classification: C91; J71; O33

3.A Introduction

Do discriminatory human decision-makers accept recommendations from non-discriminatory algorithms? Recent years have seen ambitious technical and regulatory efforts to ensure that predictive algorithms used in high-stakes domains, such as hiring and criminal sentencing, do not discriminate against protected groups. Yet in precisely these settings, humans typically retain full discretion and authority. The effectiveness of these efforts therefore depends on whether they adopt and act on these tools. Whether this is affected by fairness constraints in the algorithm design, particularly among those who themselves base decisions on group membership, is unclear. This paper presents an experiment to explore the mechanisms and implications of placing a non-discriminatory algorithm in the hands of a discriminatory human decision-maker.

According to the economics literature, human discrimination is primarily driven by statistical beliefs (Phelps, 1972; Arrow, 1973). This framework assumes that decision-makers hold prior beliefs about average group differences in an unobservable outcome of interest (e.g., job performance) (Bordalo et al., 2019; Bohren et al., 2019, 2025). In the absence of sufficient individual-level information, they use group membership to infer that outcome for a given individual (e.g., a job candidate). As individual-specific signals become available, these beliefs are updated, and group-based inference declines. In principle, algorithmic predictions should reduce statistical discrimination by providing informative, individual-level signals. However, when decision-makers retain full discretion, this effect depends on whether they follow these predictions. Fairness interventions that make algorithmic predictions uninformative with respect to group membership may reduce adoption when group membership is believed to be predictive. Behavioral economics research shows that perceived signal strength depends not only on content but also on the signal-generation process, which individuals assess through their priors (Gentzkow and Shapiro, 2006; Ambuehl and Li, 2018; Conlon et al., 2022; Chopra et al., 2024). When this process diverges from beliefs about relevant predictors, the signal may be discounted, resulting in limited belief updating and persistent statistical discrimination.

Another source of discrimination arises from individual preferences for or against certain groups, commonly referred to as taste-based discrimination (Becker, 1957). Unlike statis- tical discrimination, which stems from information asymmetries, taste-based discrimination reflects the decision-makers' underlying preferences and is therefore unlikely to be influenced by the provision of non-discriminatory algorithmic recommendations (Oreopoulos, 2011; Delavande and Zafar, 2018).

I design an online hiring experiment to study how the exclusion of group membership from the input data of algorithms[1] providing predictions about candidates' job performance affects (i) belief updating about candidates, and (ii) discrimination in final hiring decisions. The experiment uses gender as the protected attribute, and job performance is defined as a candidates relative ranking on a math and science test. Such tests are often associated with the belief that men perform better, despite no actual gender difference in outcomes (Exley and Nielsen, 2024). I elicit participants prior beliefs about gender differences in performance, that is, whether they believe gender predicts job performance, and measure how these beliefs relate to (i) belief updating, and (ii) hiring behavior. The algorithm is explicitly designed and explained as highly informative, due to its access to predictor variables that are unobservable to participants. All predictions are accurate and equal in their categorization of candidates across genders. Thus, fully following them is payoff-maximizing and implies no discrimination between male and female candidates.

The design has several advantages, particularly over field settings. First, it allows me to disentangle the underlying drivers of discrimination with non-discriminatory algorithms, separating statistical beliefs from taste-based motives. This distinction is policy-relevant: when fairness interventions are undermined by statistical beliefs, targeted information interventions

[1] Excluding protected attributes such as race or gender from algorithmic input data is a common fairness intervention, often motivated by legal concerns about disparate treatment. In the U.S., this practice is guided by the Equal Protection Clause of the Fourteenth Amendment and, in employment contexts, by Title VII of the Civil Rights Act of 1964. Predictive algorithms used in the U.S. criminal justice system, for example, omit race when assessing reoffending risk. From a technical perspective, this approach is contested, as other variables may act as proxies and reproduce disparities; see Barocas and Selbst (2016); Gillis and Spiess (2019) for discussion.

to improve the perceived accuracy of non-discriminatory algorithmic predictions may be effective. In contrast, adverse effects driven by tastes may require institutional responses, such as changes to decision-making authority. Second, I generate genuine algorithmic performance predictions using a model trained on data collected in a separate pre-study. This allows me to vary the algorithms input data while holding prediction outcomes and all other factors constant. This is an advantage over field settings, where changes to the algorithms inputs are often confounded with changes in outcomes. Third, to isolate reactions to the exclusion of gender specifically from any interventions to the input data, I implement a placebo treatment in which a non-stereotypical attribute is excluded instead.[2]

I find suggestive evidence that protecting group membership in algorithmic predictions leads participants to update their beliefs about candidates more conservatively. Contrary to expectations, this effect is not related to prior beliefs about gender as a predictor. Instead, the intervention appears to generate general uncertainty about the algorithms informativeness across participants, resulting in less correction of beliefs about candidates. Regarding behavioral effects, I find a significant increase in discrimination under the gender-blind algorithm: male candidates are 63% more likely to be hired than equally qualified female candidates in this treatment. This increase appears to be driven by prior statistical beliefs. Participants who believe that men outperform women on the test are more likely to discriminate in hiring (and thus to override the algorithms recommendation) with the gender-blind algorithm as compared to the gender-aware algorithm. This effect is not driven by participants with extremely strong priors but also appears among those with more moderate ones. Finally, about 26% of participants make hiring decisions that cannot be explained by posterior beliefs, significantly favoring female candidates.[3] This share is robust across treatments and, given the asymmetric favoring of female candidates, is likely driven by taste rather than noise. These results suggest that fairness constraints can paradoxically exacerbate human discrimination based on statistical beliefs by hindering algorithm adoption. Moreover, taste seems to remain a major source of human discrimination and, as expected, designing non-discriminatory algorithms does not seem to mitigate it.

To my knowledge, this is the first experiment to provide causal evidence on the impact of protecting group membership in algorithms on their adoption. It specifically disentangles the impact on (i) belief updating from (ii) final decision-making, and uniquely allows for measuring the association with prior statistical beliefs about protected groups. The finding that disparities in outcomes increase under the blinded algorithm broadly confirms the theoretical model by Gillis et al. (2021). They demonstrate that excluding protected characteristics from an al-

[2] This additional treatment was added to the pre-registration on September 26, 2024.

[3] This fraction is in line with findings from prior hiring experiments, e.g., Campos-Mercade and Mengel (2024).

gorithm's input leads to more discriminatory decisions by human decision-makers who hold biased beliefs about protected groups because blinded algorithms fail to sufficiently correct for these beliefs. The authors provide experimental evidence showing that when asked to predict others' math performance, participants perceive gender-specific average scores as more informative than gender-neutral ones. However, their study does not elicit participants prior beliefs about gender differences in performance, despite these beliefs being central to the theoretical framework.

While my findings replicate the predicted increase in disparities and reduced belief updating under the blind algorithm, they depart from the theoretical model in one key respect: the decline in belief updating is not moderated by prior beliefs. Instead, it appears to reflect a more general reduction in the perceived informativeness of fairness-constrained algorithms, regardless of belief accuracy. My design allows me to measure this association. Another unique feature of my design is that participants are given a genuine, highly informative algorithm that is trained on real data instead of simple group averages. This is important because it allows me to test whether excluding a single variable affects adoption even when the algorithm is known to be highly predictive.

Overall, these results suggest that human decision-makers' behavioral factors can undermine technical fairness at the system level. This is particularly relevant given the extensive resources currently devoted to designing, regulating, and implementing non-discriminatory and welfare-enhancing algorithms in high-stakes public and private sectors. However, these efforts may be ineffective if the incentives and behavior of human decision-makers are overlooked and algorithmic fairness is treated solely as a design problem rather than a problem of human-algorithm interaction.

The remainder of this chapter is structured as follows. Section 2 reviews the relevant literature. Section 3 outlines the experimental design and sample. Section 4 presents the empirical results. Section 5 discusses the findings and Section 6 concludes.

3.B Related Literature

This study contributes to four main strands of literature in economics.[4] First, it adds to the young economics literature on algorithmic fairness (Lambrecht and Tucker, 2019; Cowgill et al., 2020; Rambachan et al., 2020; Capraro et al., 2024; Lambrecht and Tucker, 2024). Much of this literature deals with identifying the sources of disparate impact, i.e., systematic differences in outcomes across protected groups, from two perspectives: (i) algorithm design, and (ii) human biases in the interaction with these tools. On the design side, prior work shows that excluding

[4] The topics discussed here are also extensively studied in other disciplines. For brevity, this section focuses on relevant work within economics.

sensitive attributes does not necessarily reduce disparities in algorithmic predictions and can even worsen them (Ludwig and Mullainathan, 2021; Kallus et al., 2022). For example,Kleinberg et al. (2018b) develop a theoretical framework and provide empirical evidence from U.S. college admissions showing that removing race from predictive models can reduce both accuracy and equity in predictions. Their model argues that fairness constraints should not be imposed at the prediction stage. Instead, predictions should be optimized for accuracy, and fairness should be introduced at the decision stage, e.g., by adjusting decision thresholds across groups. Arnold et al. (2021) find that excluding race from pretrial risk assessments does not eliminate disparities, as race is inferred from correlated inputs. White defendants are more often recommended for release than equally qualified Black defendants, even under quasi-random judge assignment. Matthew et al. (2024) show that Facebooks ad delivery system continues to produce racial disparities in targeting, and in some cases worsens them, even though race is excluded as an input. These disparities arise because the algorithm infers group membership from patterns in user behavior that serve as proxies for race.

On the behavioral side, evidence shows that decision-makers interpret algorithmic recommendations through their own beliefs and biases (Ludwig and Mullainathan, 2021; Agan et al., 2023; Pethig and Kroenung, 2023; Morewedge et al., 2023; Celiktutan et al., 2024; Glickman and Sharot, 2025). For example, Davenport (2023) finds that police officers apply risk scores selectively based on the defendants race. They use the tool less often for Black defendants in low-severity cases and issue more warrants for Black than white defendants with identical scores. This leads to racially disparate outcomes despite an explicilty neutrally designed and accurate algorithm, not because of flaws in the tool itself, but because officers apply it differently depending on the defendants race. Albright (2019) further demonstrates that judges systematically override algorithmic risk assessments in U.S. courts toward stricter bond conditions for black defendants compared to similar white defendants beyond differences in risk scores. My experimental design allows me to connect these two strands of the fairness literature: algorithm design and human behavior.

Second, this study contributes to the literature on algorithm adoption, particularly in consequential decision-making (Castelo et al., 2019; Logg et al., 2019; Burton et al., 2020; Allen and Choudhury, 2022; Garcia et al., 2024; Kim et al., 2024). Prior research indicates that adoption depends on the decision-context (e.g., task type), system-level attributes (e.g., error rates), and human decision-maker characteristics (e.g., domain expertise) (Dietvorst et al., 2015; Castelo et al., 2019; Angelova et al., 2023). Consistent with the finding that transparency and explainability foster algorithm adoption (Dietvorst et al., 2018; Kawaguchi, 2021; Reich et al., 2023), my findings indicate that participants are responsive to the specific model predictors used, even in a setting where the algorithm's accuracy remains constant and high. My findings also corroborate evidence of confirmation bias in algorithm use, as decision-makers favor predictions that align with their prior beliefs (Liu et al., 2023). Furthermore, this study contributes to the growing literature on incentive alignment between humans and machines. It demonstrates that even effective algorithms face limited adoption when the decision-maker's objectives conflict with the algorithm's goals (McLaughlin and Spiess, 2022; De-Arteaga et al., 2025). For instance, Stevenson and Doleac (2024) find in the U.S. context, that judges in the systematically override algorithmic predictions of defendants' reoffending behavior because. Specifically, they find that judges exhibit a significantly higher leniency for young defendants (often predicted to have a higher risk of reoffending) and thus diverge from the algorithms objective of minimizing reoffending risk. In line with this, my findings reveal that such preferences, independent of beliefs, may constrain algorithm adoption.

Third, this study contributes to the extensive literature on belief updating (for a review, see Benjamin (2019)). Prior research has robustly documented confirmation bias and conservatism bias in updating, i.e., the tendency to overweight prior beliefs and underreact to new information (Kahneman and Tversky, 1973; Rabin and Schrag, 1999; Mobius and Rosenblat, 2006; Weizsäcker, 2010; Coutts, 2019; Campos-Mercade and Mengel, 2024). Agarwal et al. (2023) demonstrate that these biases extend to settings with machine-generated signals: in a medical field experiment, they show that doctors do not optimally integrate AI predictions for treatment but deviate from Bayesian updating, resulting in worse outcomes than human-only or fully automated decisions. My findings also directly relate to the literature on the endogenous assessment of signal informativeness (Gennaioli and Shleifer, 2010; Enke and Zimmermann, 2019; Fryer Jr et al., 2019; Enke and Graeber, 2023), and confirms prior research showing that the source and *how* a signal is generated significantly influences its perceived strength (Griffin and Tversky, 1992; Simonsohn et al., 2008; Ambuehl and Li, 2018; Conlon et al., 2022).

Fourth, this study contributes to the extensive literature on discrimination, in particular in hiring (Goldin, 1994; Reuben et al., 2014; Bertrand and Duflo, 2017; Blau and Kahn, 2017). Recent work increasingly focuses on identifying the mechanisms underlying discriminatory behavior, in particular disentangling the role of inaccurate and accurate statistical beliefs, cognitive biases, and preferences ("taste", which can also include image concerns)(Bohren et al., 2019, 2025; Barron et al., 2024; Coffman et al., 2021b; Campos-Mercade and Mengel, 2024; Exley and Nielsen, 2024). In recent years much focus has been attributed to the impact of algorithms on discrimination in hiring, both from the perspective of candidates and managers. Avery et al. (2024) demonstrate that integrating AI into the hiring process for STEM candidates reduces the gender gap on both the demand and supply sides. Specifically, their study finds women are more inclined to apply for positions when they know they will be evaluated using AI screening. Moreover, when recruiters are informed of an applicants AI score, they exhibit a greater propensity to choose women than in situations where they are unaware of the score. Cowgill (2020) demonstrates that even when a hiring algorithm is trained with a biased data set from human decision-makers, it can discriminate less than the underlying training set. Building on this, Rabinovitch et al. (2024) show in experimental studies that individuals often fail to discount irrelevant attributes such as gender or race in hiring decisions, yet perform more accurately when receiving advice from either a human or an algorithm, suggesting that algorithmic input can help mitigate discriminatory hiring.

Most closely related to my experiment is the recent experimental study by Dargnies et al. (2024), which examines the acceptance of algorithms in hiring from the perspective of both managers and candidates. Although the algorithm outperforms the human manager, both groups initially prefer the human manager to make the hiring decision. However, when the algorithm excludes gender from its input variables, candidates prefer the algorithm to the human manager. The authors do not examine how managers respond to this intervention. My experiment addresses this gap and provides suggestive evidence that managers may be less likely to follow the recommendations of algorithms that exclude gender. As such, these results also speak to the affirmative action literature and confirm previous experimental findings highlighting the potential for reduced acceptance as a counterproductive outcome of such interventions (Holzer and Neumark, 2000; Miller and Segal, 2012; Niederle et al., 2013).

3.C Experiment

This section describes the design of the pre-studies (Section 3.C.1) and the main experiment, including details on the participant sample (Section 3.C.2). The full experimental instructions are provided in Appendix 3.A.1.

3.C.1 Pre-studies

Two pre-studies precede the main experiment, with participant recruitment conducted via Prolific in July and September 2024, respectively. In the first pre-study, 400 U.S. adults, representative of the general population, complete two standardized multiple-choice tests in math and science. This pre-study serves two purposes: (i) to generate a candidate pool for the main experiment, and (ii) to provide training data for the algorithm. One test score defines candidate performance in the main study, while the other serves as an input to the algorithm. This ensures that the algorithm produces accurate and consistent predictions across variations in the input variables. It also allows for a clear and credible explanation to participants that the algorithm is highly informative. The test questions are adapted from the Armed Services Vocational Aptitude Battery (ASVAB), following previous hiring experiments (e.g., Exley and Nielsen, 2024). Participants are given 15 seconds per question. At the end of the study, participants provide demographic information and consent to the use of their data for future research and algorithm training (for research purposes only). Average completion time is 12 minutes. Compensation includes a base payment of \$2 plus \$0.10 per correct answer.

Results show that performance on the first test is the strongest predictor of job performance, i.e. relative ranking in the main test (Pearson's $r = 0.6828$, $p < 0.001$). Given the strength of this correlation, adding demographic covariates to the prediction model does not change prediction outcomes. Among the elicited demographic variables, education, defined as holding a Bachelor's degree or higher compared to not, is significantly correlated with performance, although the correlation is modest (Pearson's $r = 0.1242$, $p = 0.0129$, 62% hold a Bachelor's degree or higher). Age, defined as being older than 45 years compared to younger adults, is also significantly correlated with performance and shows a negative relationship (Pearson's $r = -0.1602$, $p = 0.0013$,46.0% are older than 45). Gender, comparing men and women, is not a predictor of performance (Pearson's $r = 0.0072$, $p = 0.8858$, 50.7% in the top half are male). Month of birth, comparing individuals born in even-numbered months to those born in odd-numbered months (Pearson's $r = 0.0372$, $p = 0.4576$, 50.7% in the top half is born in an even-numbered month), and marital status, comparing married to unmarried participants (Pearson's $r = 0.0020$, $p = 0.9681$, 48.9% in the top half is married), are not correlated with performance.

The second pre-study elicits beliefs regarding demographic differences in performance on math and science tests. This study aims to select variables for candidate CVs and algorithm training, and confirm that beliefs about gender as a predictor of performance on a math-and-science test are sufficiently heterogeneous. In this study, I ask 300 participants, representative of the U.S. adult population, to estimate expected performance differences based on five demographic variables: gender (male vs. female), age (over vs. under 45), education (Bachelors degree or higher vs. no Bachelors degree), month of birth (even- vs. odd-numbered), and marital status (married vs. not married). Participants are informed about the test's focus on math and science, and about the fact that the sample is representative of the U.S. and thus balanced across these demographics. They are then asked to estimate the likelihood that a top performer, defined as someone whose performance ranks in the top half, belongs to each respective group.[5] Beliefs are elicited on a 0–100% scale and incentivized using the Becker–DeGroot–Marschak

[5] For example, the question used to elicit beliefs about gender is: "You are randomly assigned a top performer, defined as a participant who ranks in the top 50% of the 400 participants from the other study. What is the probability that this individual is male rather than female?"

(BDM) mechanism: if a participants belief for the selected candidate exceeds a randomly drawn number $X \in \{0, \ldots, 100\}$, they receive \$1 if the randomly selected participant belongs to the specific group, and \$0 otherwise. If their belief is below X, they receive \$1 with a probability of X% and \$0 otherwise (Karni, 2009). Participants are informed that it is in their best interest to state their true beliefs and are provided with a link to detailed information about the payoff mechanism (Danz et al., 2022). Participants receive a \$0.50 base payment and can earn an additional \$1 from one randomly selected belief.

Results show that participants, on average, believe that a top performer is more likely to have a higher level of education: the average estimated probability that a top performer holds a Bachelor's degree or higher is 72.1% (SD = 18.5). For gender, participants assign an average probability of 56.2% (SD = 16.5) to the top performer being male. For month of birth, they belief that the likelihood that a top performer was born in an odd-numbered month is 49.8% (SD = 14.3). For age, participants estimate that the average probability that a top performer is over 45 years old is 48.3% (SD = 18.2). For marital status, participants estimate that the average probability that a top performer is married is 50.7% (SD = 18.3).

Out of five demographic variables, I select three to include in the candidate CVs: education, gender, and month of birth. I include education because it is a realistic CV feature and is correctly perceived as predictive of performance (M = 72.1, SD = 18.5). Its inclusion reduces experimenter demand effects and ensures credible variation across candidates. I also include gender, which serves as the protected attribute in the main treatment. Beliefs about its predictive value exhibit sufficient heterogeneity (M = 56.2, SD = 16.5), despite the absence of actual gender differences in test performance. Moreover, gender is a realistic and widely recognized protected characteristic. Finally, I include month of birth.[6] Month of birth is perceived as unrelated to performance (M = 49.8, SD = 14.3) and is unlikely to interact with other demographic attributes when multiple variables are jointly presented. I include a placebo treatment in which month of birth, rather than gender, is excluded from the algorithms input. This allows me to disentangle whether participants respond specifically to the exclusion of gender or to the exclusion of variables more generally.

From the 400 participants in the first pre-study, I select eight individuals to serve as candidates in the main experiment. Each candidate is presented with a short CV containing binary demographic information. The eight candidates are selected to be gender balanced across education level and month of birth categories.

To train the algorithm to predict the performance on the math-and-science test (the "job task") for each of these eight workers, I use the three binary demographic variables and the score on the other math-and-science test. Specifically, I estimate a simple logistic regression

[6] The use of month of birth as a non-stereotypical attribute is adapted from Coffman et al. (2021b).

model using data from 392 of the 400 pre-study participants to predict whether or not they will be classified as "top" or "low" performers. All classifications of the eight selected candidates are accurate and balanced across the symmetric female and male candidates. Given the predictive power of the performance score on the other math-and-science test, varying the demographic inputs does not affect the classification outputs. The small number of predictors and their low correlation avoid proxy issues when excluding one of them.

3.C.2 Main study

The experiment includes a final sample of 1,251 participants (4.5% dropout rate), recruited via Prolific in September 2024. The baseline and main treatment each consist of 451 participants,[7] and the placebo treatment consists of 349 participants. Each treatment sample is representative of the U.S. adult population. Appendix Tables A4–A6 report detailed summary statistics.

The design of the main experiment closely follows the one in Campos-Mercade and Mengel (2024). Figure A1 shows the experimental stages.

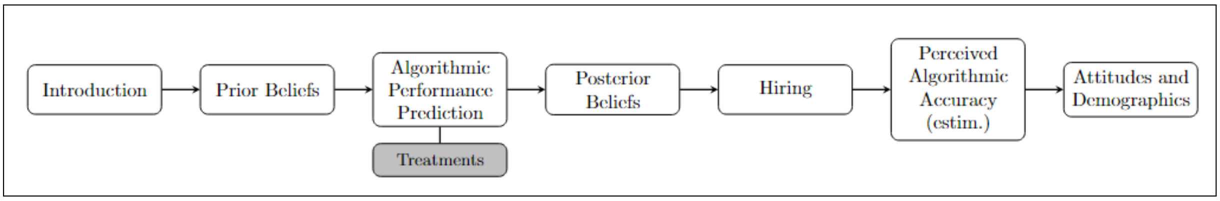
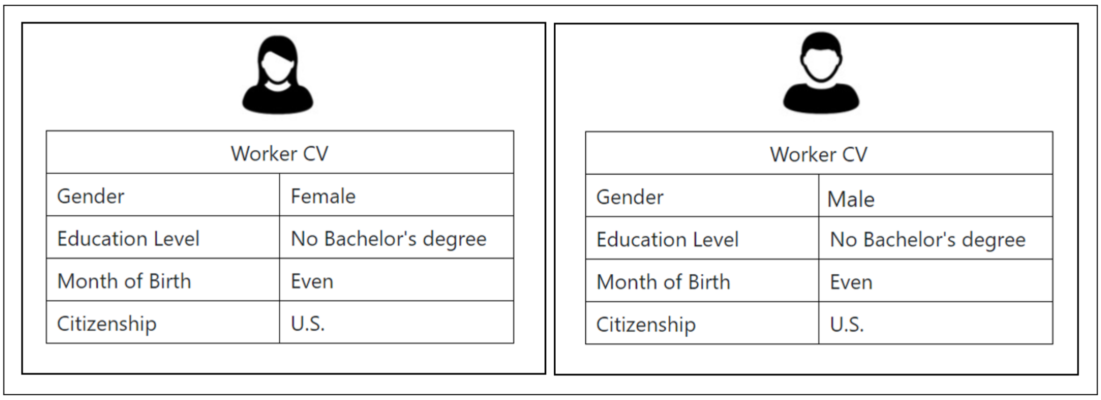

Introduction
Prior Beliefs
Algorithmic Performance Prediction
Treatments
Posterior Beliefs
Hiring
Perceived Algorithmic Accuracy (estim.)
Attitudes and Demographics

Figure A1: Stages in the Experiment

First, participants receive general instructions and are informed that 400 U.S. adults took a standardized online math and science test. They are explicitly informed that these test takers are representative of the general U.S. adult population in terms of gender balance (i.e., approximately half are female and half are male), educational level (i.e., approximately half do not have a bachelor's degree and half have a bachelor's degree or higher), and birth month (i.e., approximately half were born in an even-numbered month and half were born in an odd-numbered month). Participants are then shown eight candidate CVs in randomized order, each presented sequentially and displaying binary information on gender, education, month of birth, and U.S. citizenship. For each candidate, participants are asked to estimate the likelihood (on a scale of 0 to 100%) that s/he is a top performer, defined as someone who ranks in the top half of all 400 participants (prior beliefs). Beliefs are incentivized using the stochastic BeckerDeGrootMarschak (BDM) mechanism equal to the second prestudy (Karni, 2009): If a participants estimate exceeds a randomly drawn number $X \in \{0, \dots, 100\}$, they receive \$5 if the candidate is a top performer and \$0 otherwise. If the estimate is below X, they receive \$5 with probability $X\%$ and \$0 otherwise. Participants are informed that truthfully reporting their

[7] The identical sample sizes after attrition are coincidental.

beliefs maximizes expected payoffs and are provided with a link explaining the payment rule. This method has been shown to improve response validity and reduces experimenter demand effects, especially for stereotype-related questions (Danz et al., 2022).



Worker CV	
Gender	Female
Education Level	No Bachelor's degree
Month of Birth	Even
Citizenship	U.S.

Worker CV	
Gender	Male
Education Level	No Bachelor's degree
Month of Birth	Even
Citizenship	U.S.

Figure A2: Example CVs Presented to Participants (Female on Left, Male on Right)

After reporting their initial perceptions, participants learn that they will receive algorithmic predictions for each worker, classifying each worker as a "top" or "low" performer. They receive instructions on how the algorithm works and what data it is trained with, i.e., from the remaining 392 participants. The input variables are explicitly listed: (i) each demographic variable, and (ii) the performance score on a similar math and science test. They are explicitly explained that performance on the other test is an informative indicator of test performance. For each worker, they see their initial estimate and the algorithm's predicted category (i.e., "top" or "low" performer) and can update their prior beliefs. Participants are informed that their payoff for this estimation task is based solely on this final, updated estimate.

In the baseline condition the algorithm includes all demographic variables. In the main treatment it excludes the gender variable, while in the placebo treatment it excludes month-of-birth information. Participants in the treatment conditions are explicitly informed that the respective variable is excluded from the algorithm's training data.

After providing their final estimates (posterior beliefs), they advance to the second stage, during which they can hire any of the candidates. They make a yes/no decision for each worker, and one of these decisions is randomly selected for payment. Hiring involves risk. They are given an initial endowment of \$2.50, and their payout increases to \$5 if they hire a top performer, but decreases to \$0 if they hire someone who is not. If they do not hire a candidate and that candidate is randomly selected for payment, they keep their initial endowment. Candidates hired through the randomly selected decision receive a fixed payment, regardless of their performance. This allows for taste-based discrimination and makes the decisions consequential.

The study concludes with one incentivized question about the perceived accuracy of the algorithm and one quantitative question to measure cognitive bias, as well as two short questionnaires assessing their attitudes toward algorithms and their perceptions of gender discrimination in the US.

Throughout the study, participants must answer comprehension questions to proceed. Participants who answer incorrectly receive immediate feedback with the correct answer and a further explanation.

Participants receive a base fee of one $1 for their participation. They are compensated for one randomly selected part of the experiment, i.e., either the estimation task (beliefs) or the hiring decision. This payment structure is used to prevent participants from hedging their responses across different stages of the experiment. The average total payment is $4.80 on average, with a median completion time of 11 minutes.

3.D Results

This section reports the main results. First, I present the distribution of prior beliefs regarding gender differences in performance. Second, I study treatment differences in belief updating. Finally, I investigate ultimate outcomes in hiring discrimination across treatments. Throughout, I focus on the role of prior beliefs about gender as a predictor of performance.

3.D.1 Prior beliefs

Sufficient heterogeneity in prior beliefs about gender as a predictor of performance across the sample is a necessary condition for studying the research question. Before presenting the main results, I therefore report their distribution within the sample. Participants' initial average belief that men are top performers is 57.63%, and for women, is 56.73% [8](detailed summary statistics by treatmetns are shown in Appendix Tables A7 and A8). To quantify statistical beliefs about gender-based performance differences, i.e., gender as a perceived predictor of performance, I subtract the average values assigned to the four female candidates from those assigned to the four male candidates (Carlana, 2019). Positive values indicate a belief favoring males, negative values favoring females. The resulting distribution of these beliefs is centered around zero: 57.7% of participants report no substantial difference (defined as within ±0.5 SD on a 0–100% scale, accounting for potential noise), while 20.1% provide higher average estimates for male candidates and 22.2% for female candidates (see Appendix Figure A4).

It is unclear whether estimates that female outperform male candidates purely reflect statistical beliefs, as prior work suggests that these can be driven by image concerns and experimenter demand effects (Coffman et al., 2021b; Dargnies et al., 2024; Barron et al., 2024). I cannot

[8] This difference is smaller than expected based on results from previous experimental research Bordalo et al. (2019). One possible explanation is the inclusion of additional information about education as an informative predictor of performance.

rule out these motivations in this experiment, and the main goal of this paper is to study the role of statistical beliefs in the acceptance of non-discriminatory algorithms. I therefore exclude these observations from the following main analyses and restrict the sample to cases where the reported gender difference is non-negative ($n = 728$). Corresponding results of the full sample are documented in the Appendix. Within this restricted sample, the male-female gap in likelihood of being a top performer range from 0 to 58.75 percentage points (mean 5.7 pp., SD: 7.5). Detailed statistics can be found in Table A10.

3.D.2 Belief updating

I begin by investigating belief updating across treatment conditions. Table A1 reports the results of Ordinary Least Squares (OLS) regressions with belief updating as the dependent variable. Following the convention in prior work on Bayesian updating (Möbius et al., 2022; Campos-Mercade and Mengel, 2024), I define belief updating as the absolute change in log-odds:

$$|\log(\text{posterior odds}) - \log(\text{prior odds})| = \left| \log\left(\frac{p_i^{\text{post}}}{1 - p_i^{\text{post}}}\right) - \log\left(\frac{p_i^{\text{prior}}}{1 - p_i^{\text{prior}}}\right) \right|.$$

This transformation maps stated probabilities (p_i^{prior} and p_i^{post}, representing participant i's belief on a 0100% scale that a candidate is a top performer before and after observing the algorithmic prediction, respectively) onto an unbounded scale to ensure comparability across participants. Observations with initial estimates (prior beliefs) of 0 or 100 that are not updated are excluded, as they result in undefined log-odds transformations. I exclude observations where participants update their beliefs in the direction opposite to the algorithmic prediction (e.g., a downward update following a "top" prediction), as these are likely to reflect errors and could bias the results. Results including these cases, for both the restricted and the full sample, are reported in Table A11.

I find a marginally significant average treatment effect in Model (1): the absolute change in participants' belief log-odds (belief updating) is smaller when the algorithm excludes gender ($p = 0.071$). This result remains robust when controlling for gender difference priors (i.e., when gender is believed to be a predictor of performance) in Model 2 ($p = 0.076$) and algorithm/candidate characteristics in Model 3 ($p = 0.076$). The interaction term between gender-blind treatment and gender difference priors is not significant ($p = 0.750$ in Model 2, $p = 0.748$ in Model 3), nor is the main effect of gender difference priors ($p = 0.548$ in Model 2, $p = 0.549$ in Model 3). Contrary to expectations, reduced belief updating under the gender-blind algorithm is not moderated by prior beliefs about gender as a predictor of performance. Instead, the intervention appears to generate general uncertainty about the algorithms informativeness. This leads to weaker correction of beliefs about candidates under this algorithm.

Table A1: Belief Updating

	(1)	(2)	(3)
Gender-blind	-0.088^{*}	-0.087^{*}	-0.087^{*}
	(0.049)	(0.049)	(0.049)
Mob-blind	0.010	0.001	0.001
	(0.055)	(0.055)	(0.055)
Gender diff. priors (std.)	–	0.024	0.023
		(0.039)	(0.039)
Algorithm prediction: low	–	–	0.008
			(0.019)
Candidate gender: male	–	–	-0.020
			(0.013)
Gender-blind × Gender diff. priors (std.)	–	0.016	0.016
		(0.051)	(0.051)
Mob-blind × Gender diff. priors (std.)	–	0.046	0.046
		(0.055)	(0.055)
Observations	5,383	5,383	5,383
Controls	Yes	Yes	Yes

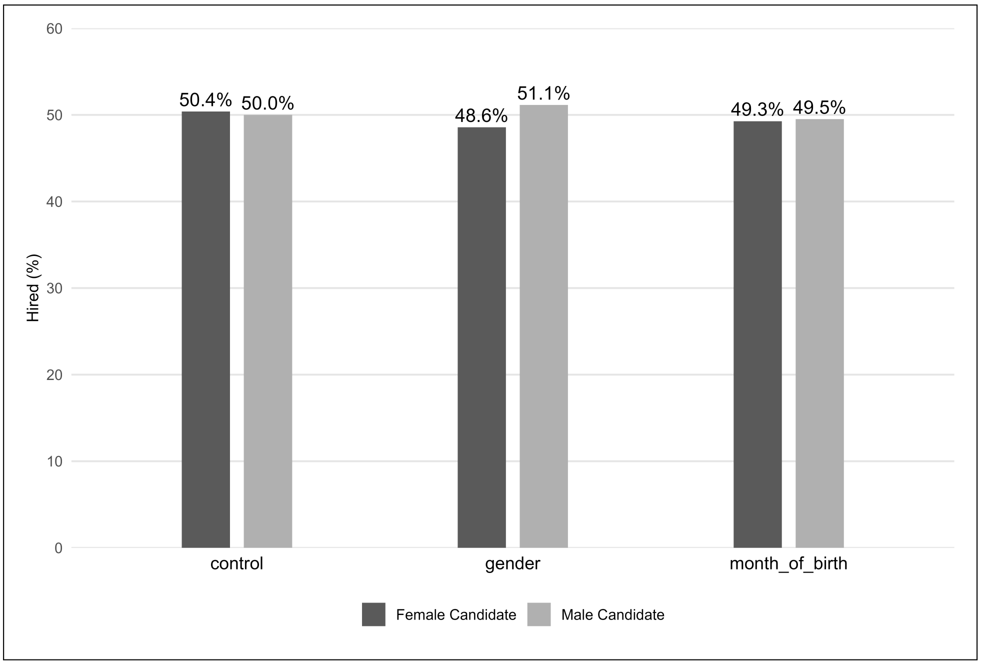
Notes: This table reports results from OLS regressions with belief updating as the dependent variable. Belief updating is defined as the absolute change in log-odds between prior and posterior beliefs, $|\log(\text{posterior odds}) - \log(\text{prior odds})|$. *Gender-blind* and *Mob-blind* indicate treatment conditions in which the algorithm excludes gender or month-of-birth information, respectively. *Gender diff. priors (std.)* measures the standardized statistical belief that males outperform females, i.e., that gender predicts performance. *Algorithm prediction: low* is a binary indicator for a low performer (vs. top performer) prediction. *Candidate gender: male* indicates the candidate's gender. All models control for participant age, gender, education, attitudes toward algorithms, cognitive bias, and priors (log-odds). The sample excludes observations where participants updated in the direction opposite to the algorithmic prediction. Standard errors are clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Reactions are also not driven by the type of signal, i.e., "top" vs. "low" prediction ($p = 0.687$), or by the gender of the candidate being evaluated ($p = 0.123$). However, the marginally negative coefficient on candidate gender may indicate greater conservatism when evaluating male candidates. These two results align with those of the hiring experiment by Campos-Mercade and Mengel (2024), who report no significant effects on belief updating based on the type of signal (in their case "positive" vs. "negative"), but a marginally significant effect for greater conservatism when evaluating male candidates. Removing month of birth information does not affect belief updating, with consistently high p-values in Model 1 ($p = 0.861$), Model 2 ($p = 0.981$), and Model 3 ($p = 0.980$). This suggests that reduced belief updating is specific to the exclusion of gender information rather than to variable exclusion in general. Full regression results, including results for all control variables and also for the full sample, are reported in Appendix Table A12.

The next section turns to the *behavioral* implications and examines how excluding gender from algorithms affects participants' discrimination in hiring.

3.D.3 Hiring

A total of 2,903 candidates are hired (49.8%), with no gender difference at the aggregate level (49.4% female, 50.3% male, $p = 0.529$, two proportion test). Figure A3 shows the aggregate hiring rates of female and male candidates by treatment. In each treatment, the average hiring rates for male and female candidates are nearly identical, with small and statistically insignificant differences (male female hiring rates: control: 0.4 percentage points, $p = 0.896$; month-of-birth: +0.2 pp, $p = 0.961$; gender-blind: +2.5 pp, $p = 0.256$). The gender gap slightly increases in the gender-blind condition compared to the control group (difference-in-differences: +2.9 pp, $p = 0.526$, χ^2 test) and the month-of-birth treatment (+2.3 pp, $p = 0.850$, χ^2 test), yet neither difference is statistically significant. The reduction in female hiring in the gender-blind treatment (48.6%) compared to the control group (50.4%) is not statistically significant ($p = 0.203$), nor is the difference relative to the month-of-birth treatment (49.3%, $p = 0.382$). Likewise, the increase in male hiring in the gender-blind treatment (51.1%) compared to the control group (50.0%) is not statistically significant ($p = 0.300$), nor is the difference relative to the month-of-birth treatment (49.5%, $p = 0.242$).



Notes: This figure shows the percentage of female and male candidates hired in each treatment condition.

Figure A3: Hiring rates by gender across treatment conditions.

At the individual level, I find significant treatment effects on discriminatory hiring behavior. Table A2 presents the results of logistic regression models, where the dependent variable is hiring discrimination (see Table A13 for full output). Discrimination is coded as 1 if a male candidate is hired while an otherwise identical female candidate is not, and 0 if either both or neither candidate is hired (Coffman et al., 2021b). I exclude cases of discrimination in favor of women (coded 1) to align with the exclusion of observations of participants holding pro-female prior beliefs and to reduce noise in results. Table A14 reports OLS results using the full sample, including cases of discrimination against male candidates.

Table A2: Discrimination in Hiring

	(1)	(2)	(3)
Gender-blind	0.490**	0.413*	0.399*
	(0.212)	(0.219)	(0.242)
Mob-blind	0.062	0.085	0.373
	(0.234)	(0.242)	(0.256)
Gender diff. priors (std.)		0.422***	0.367*
		(0.116)	(0.210)
Gender diff. priors (std.)^2			0.016
			(0.050)
Gender-blind × gender diff. priors (std.)		0.267*	0.204
		(0.156)	(0.329)
Mob-blind × gender diff. priors (std.)		−0.226	0.499
		(0.168)	(0.389)
Gender-blind × gender diff. priors (std.)^2			0.031
			(0.131)
Mob-blind × gender diff. priors (std.)^2			−0.395**
			(0.160)
Observations	2,735	2,735	2,735
Controls	Yes	Yes	Yes

Notes: This table reports results from logistic regression models with discrimination as the dependent variable. Discrimination is coded as 1 if a male candidate is hired and a female candidate with otherwise identical characteristics is not, and 0 if both or neither are hired. Observations where the female candidate is hired and the male candidate are excluded. *Gender-blind* and *Mob-blind* indicate treatment conditions in which the algorithm excluded gender or month-of-birth information, respectively. *Gender diff. priors (std.)* measures prior beliefs that male candidates outperform female candidates, i.e., that gender is a performance predictor. All regressions control for participant age, gender, education, and algorithm attitudes. Standard errors are clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Model 1 estimates the average treatment effects on discriminatory hiring. In the gender-blind condition, discrimination is significantly higher than in the control group ($\beta = 0.49$, $p = 0.021$). This corresponds to a 63% rise in the odds of hiring a male over an equivalent female candidate when the algorithm excludes gender ($\exp(0.49) \approx 1.63$). In contrast, the month-of-birth treatment shows no significant effect ($\beta = 0.06$, $p = 0.79$). These results suggest that the increase in discrimination is specific to the exclusion of gender in the algorithm, which is in line with the earlier finding on reduced belief updating under this algorithm.

Model 2 adds prior statistical beliefs about gender differences in performance to investigate potential mechanisms behind the treatment effect. The average effect of the gender-blind treatment remains significant, though only marginally ($\beta = 0.413$; $p = 0.059$), which suggests that priors partially mediate the treatment effect. The main effect of priors is highly significant ($\beta = 0.422$; $p < 0.001$). This shows that participants base their hiring decisions on their prior beliefs about gender differences, and thus tend to override the algorithmic recommendations (which are designed to be balanced across gender). Moreover, the interaction between the gender-blind condition and prior beliefs is positive and marginally significant ($\beta = 0.267$, $p = 0.089$). This implies that for each standard deviation increase in the belief that gender predicts performance, the log-odds of hiring discrimination are 0.267 higher under the gender-blind condition relative to the control. In contrast, the interaction effect for the month-of-birth condition is not statistically significant ($\beta = -0.226$, $p = 0.179$), suggesting that the observed effect is specific to the exclusion of gender rather than to algorithm intervention more generally. Taken together, these results indicate that excluding sensitive attributes from algorithms may increase human discrimination based on statistical beliefs.

Model 3 includes a squared term for standardized priors to test whether the moderating effect of beliefs is concentrated among individuals with extreme beliefs. The average treatment effect of the gender-blind condition remains marginally significant ($\beta = 0.399$, $p = 0.098$). The main effect of prior beliefs is also marginally significant ($\beta = 0.367$, $p = 0.080$). Their squared term is not significant ($\beta = 0.016$, $p = 0.746$) and neither is its interaction with the gender-blind treatment ($\beta = 0.031$, $p = 0.813$). The interaction between prior beliefs and the gender-blind condition, marginally significant in Model 2, loses significance ($\beta = 0.204$, $p = 0.536$) likely due to collinearity or reduced power. These results suggest that the increase in discrimination under the gender-blind algorithm is not limited to participants with extreme priors, but may also extend to those with more moderate beliefs. However, due to limited statistical power and potential multicollinearity, I cannot rule out that nonlinear patterns exist but remain undetected.

I find an unexpected significant interaction between the squared term of prior beliefs and the month-of-birth treatment ($\beta = -0.395$, $p = 0.014$). Among participants who strongly believe that gender predicts performance, discrimination decreases when the algorithm excludes month-of-birth information. These participants therefore follow the algorithmic recommendations more often then, likely because they perceive it as more credible when it relies only on variables they consider relevant: gender, education, and performance on another test. This result from the placebo treatment illustrates that even minor changes to the algorithm's input can influence its adoption (Dietvorst et al., 2018), which is, in turn, driven by human decision-makers' priors about the predictive value of input features.

Taste-based hiring

Hiring decisions are not driven solely by beliefs about candidate performance. I find that a total of 25.54% of all participants deviate from their posterior beliefs in their hiring behavior, i.e., they hire a candidate they believe has less than a 50% chance of being a top performer after viewing the algorithm's prediction. This fraction is consistent with previous findings in Campos-Mercade and Mengel (2024). I find no significant difference in this behavior across treatments (control: 23.0%, gender-blind: 26.7%, birth month: 27.3%; $\chi^2(2) = 1.42$, $p = 0.491$). I also find that this behavior does not vary significantly with participant demographics or education level (see Appendix Table A15), again consistent with Campos-Mercade and Mengel (2024).

The data suggest that these deviations are asymmetric across candidate gender and systematically favor female candidates. When participants assigned low probabilities (below 50%) to a male and the equivalent female candidate, 11.1% of them hired the female but not the male at least once, while 7.6% did the opposite. This difference is statistically significant ($\chi^2(1) = 4.60$, $p = 0.032$, McNemar's chi-squared test). I find this in all treatment groups (control: 10.3% female vs. 5.7% male ($p = 0.090$); gender-blind: 11.8% female vs. 6.5% male ($p = 0.061$); birth-month treatment: 12.7% female vs. 7.8% male ($p = 0.165$), note that the sample size in the birth-month treatment is smaller than in the other two treatments). While the treatment-specific differences do not reach the 5% significance level, the significant result in the pooled sample and the similar behavior across treatment groups suggests that this behavior is driven by preferences ("taste"), and not only noise. This result is consistent with Coffman et al. (2021b), who also find that hiring decisions not explained by statistical beliefs tend to favor female candidates.

At the decision level, 10.16% of all hires are inconsistent with participants posterior beliefs. Consistent with the participant-level analysis, these deviations appear asymmetric and favor female candidates. Table A3 presents results from a logistic regression using a restricted subsample of hiring decisions in which participants assigned both a male and an equivalent female candidate a posterior probability below 50% of being a top performer. The dependent variable, *discrimination*, is coded as 1 if only one of the two candidates is hired and 0 if both or neither is hired. The results indicate that female candidates are significantly more likely to be hired than male candidates in these belief-inconsistent cases ($p = 0.090$). Appendix Table A16 shows results for the full sampleincluding participants with pro-female priorsand confirms the same direction of the effect, with a smaller p-value ($p = 0.036$).

Table A3: Taste-Based Hiring (Decision-Level)

Candidate gender: female	0.1733*
	(0.1023)
Gender-blind	0.0975
	(0.2592)
Mob-blind	0.0483
	(0.2715)
Participant gender: female	0.0145
	(0.2130)
Observations	1742
Controls	Yes

Notes: This table reports results from a logistic regression. It uses a subsample of observations where participants assigned predicted performance below 50% to both an equivalent male and female candidate (posterior beliefs). Discrimination is coded as 1 if only one of the two candidates is hired, and 0 if both or neither are hired. Controls include participant age and education. Standard errors are clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Further results

Confirmation bias: I find strong evidence of confirmation bias in participants adoption of algorithmic recommendations, in line with previous research (Liu et al., 2023; Davenport, 2023). When the algorithm predicts both a lower-educated male and female candidate to be top performers, participants are significantly more likely to hire only the male ($\beta = 1.979$, $p < 0.001$, full logistic regressions results are in Table Appendix Table A17). This implies that participants appear to adopt the algorithms recommendation when it aligns with their prior beliefs, but discount it when it contradicts those. Although this paper focuses on the decision-maker's behavior (the demand side), these results also reveal implications for the supply side: female candidates with lower education are comparatively disadvantaged, even when supported by non-discriminatory and accurate algorithmic recommendations. These findings demonstrate that confirmation bias in human decision-making can undermine the intended neutrality of algorithmic decision tools.

Ingroup Bias: Participants exhibit ingroup bias in their hiring behavior, consistent with Coffman et al. (2021b). Appendix Table A13 and Table A14 report full regression outputs, including demographics, using the restricted and full samples, respectively. In both samples, female participants are less likely than male participants to discriminate against female candidates. In the restricted sample, which excludes participants who believe that women outperform men, this difference is large and marginally significant ($\beta = -0.304$, $p = 0.087$). When controlling for prior beliefs, the effect diminishes and becomes statistically insignificant ($\beta = -0.234$, $p = 0.176$). This suggests that beliefs explain part of the gender difference. In the full sample, the gender gap remains large and statistically significant across all models, even after controlling for priors: $\beta = -0.038$ ($p = 0.002$) in Model 1, $\beta = -0.037$ ($p = 0.002$) in Model 2, and $\beta = -0.035$ ($p = 0.003$) in Model 3.

Perceived algorithm accuracy: I find significant treatment effects in perceived algorithm accuracy. Participants rate the gender-blind algorithm as less accurate in predicting candidate performance (mean accuracy: 62.4%) than both the control algorithm that uses all available variables (mean accuracy: 65.6%, $p = 0.0017$, one-sided t-test) and the algorithm that excludes month-of-birth information (mean accuracy: 64.7%, $p = 0.027$, one-sided t-test). Regression results in Appendix Table A18 confirm this finding and show that the reduction in perceived accuracy under the gender-blind treatment is not moderated by prior beliefs about gender as a predictor of performance. This aligns with the earlier result that belief updating is more conservative under the gender-blind algorithm, regardless of participants prior beliefs.

3.E Discussion

Discussions of algorithmic fairness interventions often adopt a technical or normative lens. This paper takes a behavioral approach and presents evidence from a hiring experiment on how fairness-constrained algorithms interact with human decision-makers. I find suggestive evidence that protecting group membership in algorithmic predictions makes individuals more conservative in updating their beliefs about candidates, regardless of their prior beliefs about group differences. This is supported by the finding that "blind" algorithms are also perceived as significantly less accurate, regardless of participants' priors. When it comes to hiring decisions, I find that discrimination in hiring increases under the fairness intervention, driven by those who initially believe that group membership predicts performance. The findings also highlight a distinct source of human discrimination that lies outside the scope of algorithmic design: individual preferences ("taste") for or against certain groups.

These results are novel in the algorithmic fairness debate, yet align with well-established behavioral findings from other domains. A large body of work on affirmative action policies shows that well-intentioned fairness interventions can elicit unintended responses by reducing trust in the decision-making process, generating greater resistance from decision makers when their goals are misaligned with the goals of the affirmative action policy, and ultimately even exacerbating disparities in outcomes (Holzer and Neumark, 2000; Miller and Segal, 2012; Niederle et al., 2013). Although the elimination of the gender variable in this setting was not explicitly framed as a fairness intervention, participants may likely have perceived it as such (Anderson et al., 2006; Fryer et al., 2008; Hughey, 2022). These findings are also consistent with prior research on algorithm acceptance and trust (Dietvorst et al., 2018; Burton et al., 2020). People are highly sensitive to how algorithms are designed and framed when evaluating their credibility . If users do not understand or trust an algorithms design, they are less likely to follow it. Finally, cognitive biases well, documented in the literature, may also help explain these effects (Benjamin, 2019; Exley and Nielsen, 2024). In this experiment, participants respond disproportionately to the removal of a single input variable, even though they know that the algorithm includes other strong predictors. Also, they do not fully update their beliefs in response to highly informative signals (even those who estimate them to be accurate), which is consistent with the large literature on errors in Bayesian updating.

Importantly, these results should not be interpreted as a case against anti-discrimination approaches at the system-level or the critical role of human oversight in high-stakes domains. In fact, both are core components of recent AI regulation, such as the EU AI Act (2024), and are individually essential for responsible AI: the former for preserving equity and mitigating discriminatory development and outcomes, and the latter for ensuring accountability, risk minimization, and fostering institutional compliance. Rather than challenging these principles, the findings underscore the need to design institutional frameworks that account for how these components interact in practice. This includes defining the optimal delegation of decision-making authority between algorithms and humans (Athey et al., 2020; Garcia et al., 2024), particularly when their objectives are not fully aligned (Cowgill and Tucker, 2020; McLaughlin and Spiess, 2022). It also complicates the question of transparency and explainability in decision procedures, both in terms of how the algorithm operates and why human decision-makers override it (Angelova et al., 2023). Finally, it raises the issue of where fairness interventions should be applied: at the prediction stage, where models are trained, or at the application stage, where decisions are made (Kleinberg et al., 2018b).

The results of this study should be interpreted with some caution due to several limitations. The experiment relies on a highly stylized setting in which the algorithm is assumed to be correct, informative, and non-discriminatoryboth in treatment and in impact. These assumptions are necessary to ensure internal validity and to isolate the behavioral response to fairness constraints. However, they likely come at the cost of reduced external validity. Algorithms used in practice often do not outperform human judgment and involve trade-offs between different fairness notions (Narayanan and Kapoor, 2024). This study does not claim that algorithmic predictions are inherently more accurate or less discriminatory than humans in real-world applications. Another important limitation concerns the relatively low heterogeneity in participants' endogenous prior beliefs about group differences, which is lower than in comparable experimental studies. In addition, a large share of participants believed that females outperform males, which led to the exclusion of more observations than originally anticipated.

This reduces the statistical power to detect interactions between prior beliefs and treatment effects and suggests that future research may benefit from approaches that exogenously create greater variation in prior beliefs. Furthermore, the experimental design is intentionally stylized and uses a simplified fairness intervention, i.e., the exclusion of a single input variable, which does not reflect how fairness is operationalized in most real-world algorithmic systems. Thus, the results may not generalize to settings with more complex algorithmic fairness approaches. Finally, in this stylized setting, decision makers are assumed to be motivated solely by accuracy in predicting candidate performance, which is implicit in the incentive structure of the experiment. In real-world contexts, however, decision makers often pursue multiple objectives that may shape their response to algorithmic recommendations. For example, managers who value team diversity may be more likely to follow fairness-oriented algorithmic hiring tools.

In addition to addressing these limitations, also through the use of observational data from field settings where fairness-aware algorithms are increasingly adopted, future research could explore potential remedies to the backfiring behavioral effects identified in this study. One avenue may be to examine information interventions that increase trust in the algorithms accuracy, provided the tool is well-designed (Haaland et al., 2023). This could be particularly relevant, as low perceived accuracy appears to hinder adoption. Another promising direction is to shift from individual to collective decision-making, which has been shown to mitigate the influence of individual biases and beliefs (Mann, 2020). Finally, building on Esponda et al. (2023), future work could investigate whether disclosing sensitive attributes only after the algorithmic prediction has been presented increases the acceptance of non-discriminatory algorithms and thereby helps reduce statistical discrimination.

3.F Conclusion

Efforts to reduce discrimination when algorithms are used in consequential decision-making may fall short if the behavioral dynamics of humanalgorithm interactions are not considered. Even when provided with non-discriminatory and accurate algorithmic recommendations, human decision-makers may continue to rely on their own statistical beliefs or taste-based preferences. As a result, disparities in outcomes may persist unless such behavioral responses are explicitly addressed. This confirms prior calls to extend the focus of algorithmic fairness beyond model design to the broader institutional context in which these tools are deployed (Mitchell et al., 2021; Barocas et al., 2023; Corbett-Davies et al., 2023). This is particularly important given the potential welfare and equity gains of algorithms used in domains such as hiring, lending, or criminal justice (Kleinberg et al., 2018b,a), and the substantial resources currently invested in their development and deployment. Realizing the intended social impact of fairness interventions requires interdisciplinary approaches that integrate technical, behavioral, and institutional perspectives.

3.A Appendix

Table A4: Gender Distribution of the Sample

Gender	Count	Percentage (%)
Female	633	50.60
Male	602	48.12
Non-binary	15	1.2
Prefer not to say	1	0.08
Total	1251	100

Notes: This table displays the gender distribution of the sample.

Table A5: Age Distribution of the Sample

Age Category	Count	Percentage (%)
18-29	231	18.47
30-39	266	21.26
40-49	181	14.47
50-59	253	20.22
60-69	238	19.02
70-79	75	6.00
>80	7	0.56
Total	1251	100

Notes: This table displays the age distribution of the sample.

Table A6: Education Levels Distribution of the Sample

Education	Count	Percentage (%)
Bachelor's degree	419	33.49
Some college credit, no degree	264	21.1
Master's degree	170	13.59
High school graduate (or equivalent)	149	11.91
Associate degree	135	10.79
Trade/technical/vocational training	47	3.76
Doctorate degree	24	1.92
Professiol degree	24	1.92
Some high school, no diploma	13	1.04
Prefer not to say	4	0.32
No schooling completed	1	0.08
Nursery school to 8th grade	1	0.08
Total	1251	100

Note: The table reports the distribution of education levels in the sample.

Table A7: Summary Statistics of Prior Beliefs About Male candidates (by Treatment)

Treatment	Mean	SD	Min	Q1	Median	Q3	Max	N
Control	57.9	10.2	27.2	50.6	57.0	64.0	90.0	451
Gender	56.8	9.29	25.0	50.1	56.2	62.5	84.8	451
Month-of-Birth	58.3	9.86	23.8	50.8	57.5	65.0	82.8	349

Notes: This table reports summary statistics of participants' prior beliefs about male candidates performance, measured as the estimated probability (0100) that candidate is a top performer. Results are shown separately by treatment group.

Table A8: Summary Statistics of Prior Beliefs About Female candidates (by Treatment)

Treatment	Mean	SD	Min	Q1	Median	Q3	Max	N
Control	57.5	10.3	26.2	50.2	56.5	63.8	100.0	451
Gender	56.0	10.0	10.5	50.0	55.0	62.2	85.0	451
Month-of-Birth	56.7	10.6	25.0	50.0	55.5	63.8	84.2	349

Notes: This table reports summary statistics of participants' prior beliefs about female candidates performance, measured as the estimated probability (0100) that candidate is a top performer. Results are shown separately by treatment group.

Table A9: Summary Statistics of Prior Beliefs about Gender Differences (by Treatment)

Treatment	Mean	SD	Min	Q1	Median	Q3	Max	N
Control	0.4	8.91	-50.0	-3.1	0.0	3.8	50.2	451
Gender	0.8	8.74	-27.8	-2.5	0.0	3.8	58.8	451
Month-of-Birth	1.6	9.03	-29.0	-2.5	0.0	4.5	37.2	349

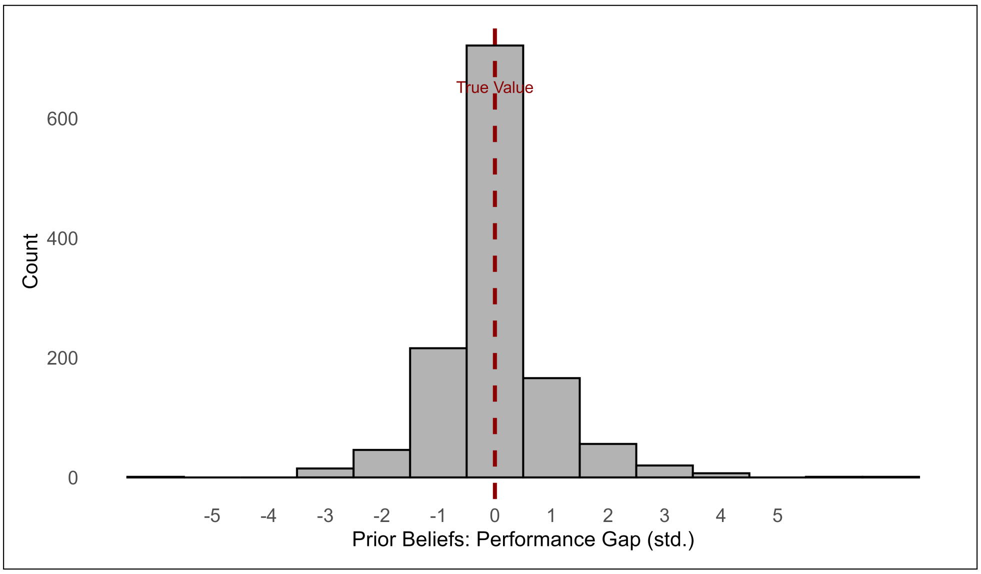
Notes: This table reports summary statistics of the difference in participants' prior beliefs about male and female candidate performance, measured as the estimated probability (0-100) that the candidate is a top performer. Positive values indicate a belief that male candidates are more likely to be top performers. Results are shown separately by treatment group.

Table A10: Summary Statistics of Prior Beliefs about Gender Differences (Subsample)

Treatment	Mean	SD	Min	Q1	Median	Q3	Max	N
Control	5.3	6.96	0.0	0.0	2.5	8.0	50.2	261
Gender	5.5	7.60	0.0	0.2	2.8	7.0	58.8	262
Month-of-Birth	6.4	8.02	0.0	0.2	3.2	9.2	37.2	205

Notes: This table reports summary statistics of the difference in participants' prior beliefs about male and female candidate performance, measured as the estimated probability (0-100) that the candidate is a top performer. Observations with pro-female prior beliefs are excluded. Results are shown separately by treatment group.

Figure A4: Distribution of Prior Beliefs About Gender Performance Gap (standardized)



Notes: This figure displays the distribution of participants' standardized prior beliefs about the performance gap between male and female candidates (male − female). The performance gap is measured as the average estimated probability that a candidate is a top performer for male candidates minus that for female candidates. The x-axis is centered at zero, with each bar representing a one standard deviation interval. Zero indicates no perceived performance difference, meaning the belief is that gender is not a performance predictor. Positive values indicate that a belief that males perform better, and negative values indicate a belief that females perform better.

Table A11: Belief Updating (Incl. Updating in Opposite Direction)

	Restricted Sample	Full Sample
Gender-blind	-0.088^{*}	-0.042
	(0.049)	(0.037)
Mob-blind	0.010	0.028
	(0.055)	(0.042)
Gender diff. priors (std.)	0.023	0.024
	(0.039)	(0.029)
Algorithm prediction: low	0.008	0.016
	(0.019)	(0.015)
Candidate gender: male	-0.020	-0.006
	(0.013)	(0.011)
Gender-blind × gender diff. priors (std.)	0.016	-0.031
	(0.051)	(0.041)
Mob-blind × gender diff. priors (std.)	0.046	0.029
	(0.055)	(0.042)
Prior log-odds	0.019	0.018^{*}
	(0.013)	(0.010)
Age	-0.006^{***}	-0.005^{***}
	(0.001)	(0.001)
Gender: Non-binary	0.062	0.119
	(0.194)	(0.113)
Gender: Prefer not to say	0.159	-0.068
	(0.202)	(0.202)
Gender: Female	0.059	0.062
	(0.045)	(0.033)
Education: Bachelor's	0.036	-0.014
	(0.083)	(0.064)
Education: Doctorate	-0.060	-0.125
	(0.131)	(0.107)
Education: High school	-0.123	-0.056
	(0.096)	(0.073)
Education: Master's	-0.049	-0.057
	(0.089)	(0.069)
Education: Prefer not to say	-0.485^{**}	-0.352^{*}
	(0.206)	(0.206)
Education: Professional	0.129	0.013
	(0.129)	(0.096)
Education: Some college	-0.080	-0.080
	(0.088)	(0.065)
Education: Some high school	-0.055	0.045
	(0.161)	(0.133)
Education: Vocational	-0.159	-0.121
	(0.129)	(0.091)
Attitude toward algorithms	0.175^{***}	0.170^{***}
	(0.033)	(0.026)
Cognitive bias	-0.077^{*}	-0.026
	(0.047)	(0.038)
Observations	5,611	9,660

Notes: This table reports the full output from OLS regressions with belief updating as the dependent variable, including observations of participants that updated their estimates in the direction opposite to the algorithm. Model 1 uses the restricted sample (no pro-female prior beliefs); Model 2 uses the full sample (incl. participants with pro-female prior beliefs). Belief updating is defined as the absolute change in log-odds between prior and posterior beliefs, $|\log(\text{posterior odds}) - \log(\text{prior odds})|$. Standard errors are clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A12: Belief Updating (Full Output)

	Restricted Sample	Full Sample
Gender-blind	−0.088*	−0.042
	(0.049)	(0.037)
Mob-blind	0.010	0.028
	(0.055)	(0.042)
Gender diff. priors (std.)	0.023	0.024
	(0.039)	(0.029)
Algorithm prediction: low	0.008	0.016
	(0.019)	(0.015)
Candidate gender: male	-0.020	−0.006
	(0.013)	(0.011)
Gender-blind × gender diff. priors (std.)	0.016	−0.031
	(0.051)	(0.041)
Mob-blind × gender diff. priors (std.)	0.046	0.029
	(0.055)	(0.042)
Prior log-odds	0.019	0.018*
	(0.013)	(0.010)
Age	−0.006***	−0.005***
	(0.001)	(0.001)
Gender: Non-binary	0.062	0.116
	(0.194)	(0.113)
Gender: Prefer not to say	0.159	−0.065
	(0.202)	(0.203)
Gender: Female	0.059	0.066
	(0.045)	(0.034)
Education: Bachelor's	0.036	−0.012
	(0.083)	(0.064)
Education: Doctorate	−0.060	−0.126
	(0.131)	(0.107)
Education: High school	−0.123	−0.051
	(0.096)	(0.073)
Education: Master's	−0.049	−0.055
	(0.089)	(0.069)
Education: Prefer not to say	−0.485**	-0.344
	(0.206)	(0.207)
Education: Professional	0.129	0.011
	(0.129)	(0.096)
Education: Some college	−0.080	−0.077
	(0.088)	(0.065)
Education: Some high school	−0.055	0.051
	(0.161)	(0.133)
Education: Vocational	−0.159	−0.119
	(0.129)	(0.091)
Attitude toward algorithms	0.175***	0.171***
	(0.033)	(0.026)
Cognitive bias	−0.077*	−0.026
	(0.047)	(0.038)
Observations	5,383	9,237

Notes: Full output from OLS regressions with belief updating as the dependent variable, corresponding to Table A1. Model 1 uses the restricted sample (no pro-female prior beliefs); Model 2 uses the full sample (incl. participants with pro-female prior beliefs). Belief updating is defined as the absolute change in log-odds between prior and posterior beliefs. The sample excludes observations where participants update in the direction opposite to the algorithm prediction. SE clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A13: Discrimination in Hiring (Full Output)

	(1)	(2)	(3)
Gender-blind	0.490**	0.413*	0.399*
	(0.212)	(0.219)	(0.242)
Mob-blind	0.062	0.085	0.373
	(0.234)	(0.242)	(0.256)
Gender diff. priors (std.)		0.422***	0.367*
		(0.116)	(0.210)
Gender diff. priors (std.)^2			0.016
			(0.050)
Gender-blind × gender diff. priors (std.)		0.267*	0.204
		(0.156)	(0.329)
Mob-blind × gender diff. priors (std.)		−0.226	0.499
		(0.168)	(0.389)
Gender-blind × Gender diff. priors (std.)^2			0.031
			(0.131)
Mob-blind × Gender diff. priors (std.)^2			−0.395**
			(0.160)
Attitude toward algorithm	−0.037	0.071	0.074
	(0.155)	(0.138)	(0.139)
Age	0.009	0.005	0.005
	(0.006)	(0.006)	(0.006)
Gender: Non-binary	−13.850***	−13.614***	−13.670***
	(0.189)	(0.198)	(0.205)
Gender: Prefer not to say	−15.806***	−15.358***	−15.577***
	(0.796)	(0.917)	(0.749)
Gender: Woman	−0.299*	−0.228	−0.199
	(0.179)	(0.175)	(0.175)
Education: Bachelor's degree	−0.443	−0.403	−0.409
	(0.325)	(0.316)	(0.321)
Education: Doctorate degree	−0.035	0.363	0.413
	(0.669)	(0.659)	(0.645)
Education: High school graduate	−0.134	0.151	0.196
	(0.371)	(0.373)	(0.377)
Education: Master's degree	−0.072	0.246	0.184
	(0.352)	(0.353)	(0.350)
Education: Prefer not to say	1.006	1.485	1.636*
	(0.796)	(0.922)	(0.770)
Education: Professional degree	−1.602	−1.299	−1.295
	(1.022)	(0.987)	(0.993)
Education: Some college credit	−0.220	−0.087	−0.078
	(0.342)	(0.336)	(0.342)
Education: Some high school	0.991	1.419*	1.482*
	(0.777)	(0.793)	(0.801)
Education: Trade/technical training	0.415	0.683	0.678
	(0.436)	(0.435)	(0.432)
Observations	2,735	2,735	2,735

Notes: This table reports the full results of logistic regression models with hiring discrimination as the dependent variable, corresponding to Table A2. Discrimination is coded as 1 if a male candidate is hired while an otherwise identical female candidate is not, and 0 if both or neither are hired. Observations where the female candidate is hired but the male is not are excluded. Standard errors are clustered at the participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A14: Discrimination in Hiring (Full Sample)

	(1)	(2)	(3)
Gender-blind	0.008	0.007	0.003
	(0.014)	(0.014)	(0.014)
Mob-blind	−0.014	−0.019	−0.004
	(0.014)	(0.014)	(0.015)
Gender diff. priors (std.)		0.027^{**}	0.026^{*}
		(0.012)	(0.012)
Gender diff. priors (std.)^2			0.003
			(0.005)
Gender-blind × gender diff. priors (std.)		0.053^{**}	0.045^{**}
		(0.019)	(0.019)
Mob-blind × gender diff. priors (std.)		0.007	0.019
		(0.016)	(0.016)
Gender-blind × gender diff. priors (std.)^2			0.005
			(0.007)
Mob-blind × gender diff. priors (std.)^2			-0.015^{**}
			(0.007)
Attitude toward algorithms	0.008	0.011	0.012
	(0.009)	(0.009)	(0.009)
Age	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Gender: non-binary	-0.134^{***}	-0.130^{***}	-0.126^{**}
	(0.047)	(0.049)	(0.050)
Gender: prefer not to say	0.199	0.199^{*}	0.209^{*}
	(0.131)	(0.120)	(0.120)
Gender: Female	-0.037^{**}	-0.036^{**}	-0.034^{**}
	(0.012)	(0.012)	(0.012)
Education: Bachelor's	−0.009	−0.015	−0.014
	(0.023)	(0.021)	(0.021)
Education: Doctorate	−0.046	−0.031	−0.029
	(0.053)	(0.051)	(0.050)
Education: High school	−0.020	−0.015	−0.011
	(0.027)	(0.026)	(0.026)
Education: Master's	−0.001	0.004	0.005
	(0.025)	(0.024)	(0.024)
Education: Prefer not to say	-0.233^{*}	-0.221^{*}	-0.220^{*}
	(0.131)	(0.120)	(0.121)
Education: Professional	0.006	0.009	0.012
	(0.029)	(0.028)	(0.028)
Education: Some college	−0.004	−0.002	−0.000
	(0.024)	(0.023)	(0.022)
Education: Some high school	0.062	0.067	0.069
	(0.092)	(0.092)	(0.092)
Education: Vocational	0.011	0.014	0.016
	(0.043)	(0.041)	(0.041)
Observations	5,004	5,004	5,004

Notes: This table reports results from OLS regressions with the full sample and discrimination as the dependent variable. Discrimination is coded as 1 if a male candidate is hired and a female candidate with otherwise identical characteristics is not, −1 if the female candidate is hired and the male is not, and 0 if both or neither are hired. *Gender-blind* and *Mob-blind* indicate treatment conditions in which the algorithm excludes gender or month-of-birth information, respectively. *Gender diff. priors (std.)* measures prior beliefs that male candidates outperform female candidates, i.e., that gender is a performance predictor. Standard errors clustered at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A15: Taste-based Hiring (Participant-Level)

	(1) Restricted Sample	(2) Full Sample
Gender: Female	-0.280	-0.304^{**}
	(0.173)	(0.128)
Gender: Non-binary	-0.533	0.180
	(1.101)	(0.563)
Gender: Prefer not to say	-0.532	-0.216
	(1081.135)	(1018.000)
Education: Bachelor's degree	0.102	-0.073
	(0.308)	(0.221)
Education: Doctorate degree	0.516	0.501
	(0.626)	(0.458)
Education: High school graduate (GED)	0.242	-0.007
	(0.375)	(0.264)
Education: Master's degree	0.105	-0.044
	(0.351)	(0.256)
Education: Prefer not to say (education)	-13.230	-13.660^{*}
	(624.192)	(508.000)
Education: Professional degree	-0.729	-1.113^{*}
	(0.812)	(0.648)
Education: Some college, no degree	-0.083	-0.086
	(0.341)	(0.237)
Education: Some high school, no diploma	-0.083	-0.309
	(0.855)	(0.688)
Education: Trade/technical training	0.079	0.210
	(0.519)	(0.363)
Age	0.007	0.001
	(0.006)	(0.004)
Treatment: Gender	0.198	0.089
	(0.207)	(0.150)
Treatment: Month-of-Birth	0.237	0.148
	(0.218)	(0.159)
Observations	728	1251

Notes: This table reports results from a logistic regression. The dependent variable is a binary indicator (1 = participant hired at least one candidate rated with a posterior probability of being a top performer below 50%; 0 = otherwise). Standard errors are in parentheses. Model (1) includes restricted sample (without pro-female prior beliefs); Model (2) includes all observations. The reference category for gender is *male*, and for education levels is *no formal education.* $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A16: Taste-Based Hiring (Decision-Level, Full Sample)

Candidate gender: female	0.1662^{*}
	(0.0795)
Treatment: gender-blind	-0.0400
	(0.1842)
Treatment: month-of-birth-blind	0.0753
	(0.1931)
Participant gender: woman	-0.2180
	(0.1574)
Observations	1742
Controls	Yes

Notes: This table reports the results of a logistic regression for the full sample (including those participants who initially stated that female workers outperform male workers, i.e., pro-female prior beliefs). It uses a subsample of observations in which participants assigned a predicted performance below 50% to both an equivalent male and female candidate (posterior beliefs). Discrimination is coded as 1 if only one of the two candidates is hired, and 0 if both or neither is hired. Controls include participant age and education. Standard errors are clustered at the participant level. ${}^{*}p < 0.10$, ${}^{**}p < 0.05$, ${}^{***}p < 0.01$.

Table A17: Confirmation Bias in Algorithm Adoption

	(1)	(2)
	Restricted Sample	Full Sample
Candidate: no bachelor	-1.202^{***}	-0.015
	(0.217)	(0.014)
Algorithm: top	-0.906^{***}	0.013
	(0.198)	(0.013)
Candidate: no bachelor × algorithm: top	2.034^{***}	0.007
	(0.294)	(0.020)
Observations	2,735	5,004
Controls	Yes	Yes

Notes: Model (1) is a logistic regression. The dependent variable is discrimination, coded as 1 if the male candidate is hired and the otherwise identical female candidate is not, and 0 if both receive the same hiring outcome (either both hired or neither hired). The sample excludes cases of pro-female discrimination and participants with prior beliefs favoring female candidates. Model (2) is an OLS regression on the full sample. The dependent variable is coded as 1 if only the male is hired, -1 if only the female is hired, and 0 if both receive the same outcome. *Candidate: no bachelor* is 1 if the evaluated candidate has no bachelor's degree. *Algorithm: top* is 1 if the algorithm predicted the candidate to be a top performer. All models include controls for participant age, gender, education, and algorithm attitudes. Standard errors are clustered at the participant level. ${}^{*}p < 0.10$, ${}^{**}p < 0.05$, ${}^{***}p < 0.01$.

Table A18: Perceived Algorithm Accuracy

	(1)	(2)	(3)	(4)
Gender-blind	-3.30^{**}	-3.39^{**}	-3.16^{***}	-3.21^{***}
	(1.39)	(1.39)	(1.07)	(1.07)
Mob-blind	−1.92	−1.90	−0.95	−0.99
	(1.48)	(1.49)	(1.14)	(1.15)
Gender diff. priors (std.)		−0.45		0.89
		(0.59)		(0.75)
Gender-blind × gender priors (std.)		−2.34		−0.91
		(1.44)		(1.08)
Mob-blind × gender priors (std.)		−1.74		−0.92
		(1.49)		(1.13)
Constant	66.43^{***}	66.47^{***}	65.60^{***}	65.65^{***}
	(0.98)	(0.98)	(0.75)	(0.76)
Observations	728	728	1,251	1,251

Notes: OLS regressions with perceived algorithm accuracy as the dependent variable (measured on a 0–100 scale). *Gender-blind* and *Mob-blind* indicate treatments relative to the control. *Gender diff. priors (std.)* is the standardized belief that gender predicts performance. Models (1) and (2) use the restricted sample (excluding participants with pro-female priors). Models (3) and (4) use the full sample. Standard errors clustered at participant level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.A.1 Experimental instructions

Pre-study: Beliefs about demographic differnces in test performance)

Slide 1
Welcome! Thank you for participating in this study. The study is conducted by a researcher at the University of Hamburg, Germany.

Your Task: In a previous online study, 400 U.S. adultsrepresentative of the U.S. adult populationtook a math and science test. Your task is to estimate whether there are demographic differences in average test performance. You will be asked to make a total of five predictions.

Your Payment: You will receive a fixed payment of $0.50 for completing the study. In addition, you can earn a bonus of $1 based on the accuracy of your predictions. At the end of the study, a computer will randomly select one of your five predictions. You will receive a bonus of $1 if that prediction is accurate, or $0 if it is not. The more accurate your predictions overall, the more likely you are to receive the bonus. There are no penalties for incorrect guesses. Click here to see the exact method used to determine the bonus payout.

Confidentiality: Your answers and data will be kept anonymous and confidential at all times.

Duration and Voluntary Participation: The study takes approximately 3 minutes to complete. Participation is voluntary, and you may withdraw at any time without penalty. Compensation is only provided for completed studies.If you agree to participate, click Agree to begin, or Disagree to exit.

- Agree
- Disagree

Optional slide: Payment rule
If your prediction exceeds a randomly drawn number X between 0 and 100, then you get paid $5 if the worker you assessed is a top performer and $0 otherwise. If your prediction is below or equal to the randomly drawn number X, then you get $5 with probability X% and $0 otherwise.

Slide 2
400 U.S. adults, demographically representative of the U.S. adult population, participated in an online math and science test. The test consisted of 20 multiple-choice questions covering general science, math knowledge, arithmetic reasoning, object assembly, and mechanical understanding.
Your are now asked to estimate whether there are demographic differences in average test performance. The more accurate you are, the more likely you receive the bonus payment.

Slides 3 through 7 (random order)
400 U.S. adults, demographically representative of the U.S. adult population, participated in an online math and science test. The test consisted of 20 multiple-choice questions covering general science, math knowledge, arithmetic reasoning, object assembly, and mechanical understanding.
Do you think average performance differs by **gender**?
There are an equal number of men and women among the 400 test takers.
Now, focusing on the top performers, these are those who scored in the top 50%: If one of these top performers is drawn at random, what is the percentage chance that this top performer is male compared to female?

Likert-Scale: 0% (women perform better) 100% (men perform better)

The same question format was repeated for four additional demographic factors. The order of the five slides was randomized across participants:

- **Age:** About half of the 400 test takers were older than 45, and half were 45 or younger.
  Now, focusing on the top performers those who scored in the top 50%: If one of these top performers is drawn at random, what is the percentage chance that this top performer is older than 45 compared to being 45 or younger?

- **Month of birth:** About half of the 400 test takers were born in an odd month (January, March, May, July, September, November) and the other half were born in an even month (February, April, June, August, October, December).
  Now, focusing on the top performers those who scored in the top 50%: If one of these top performers is drawn at random, what is the percentage chance that this top performer is born in an odd month compared to being born in an even month?
- **Marital status:** About half of the 400 test takers were married and half were not married.
  Now, focusing on the top performers those who scored in the top 50% if one of these top performers is drawn at random, what is the percentage chance that this top performer would be married compared to being unmarried?
- **Education level:** Almost half of the 400 test takers had a Bachelors degree or an even higher degree. Half of them had a lower degree or no degree at all.
  Now, focusing on the top performers those who scored in the top 50% if one of these top performers is drawn at random, what is the percentage chance that this top performer holds a Bachelors degree or higher compared to not having a Bachelors degree?

Main Study

Slide 1
Welcome! Thank you for participating in this study. The study is conducted by a researcher at the University of Hamburg, Germany.
The study will take approximately 10 minutes to complete.

- Base payment: You will receive $1 for completing this study.
- Bonus payment: You can earn an additional payment of up to $6.

Incorrect answers will not be penalized and will not reduce your payment.

You will receive the base payment within 24 hours and the bonus payment within 48 hours.

Please note that compensation can only be issued if you complete the study.
Your data will remain anonymous and confidential at all times.
Your participation is voluntary, and you may withdraw at any time without penalty.
Please confirm that you have read and understood these instructions and select Confirm to continue or Decline if you choose not to participate.

- Confirm
- Decline

Slide 2
Instructions: Please read the instructions in this study carefully. You will be asked to answer comprehension questions throughout the study. There are two parts:

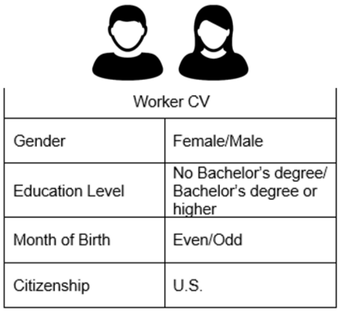
- Part 1 takes approximately 7 minutes to complete.
- Part 2 takes approximately 3 minutes to complete.

After you complete Part 1, we will provide details for Part 2. The study concludes with a brief survey.
Bonus Payments: At the end of the study, a computer will randomly select which part will count for the bonus payment. Depending on your answers, you can earn an **additional $5 in each part**. There is one additional question for which you can earn an **additional $1**.Upon completion of the study, you will be notified of your bonus payment and total payment.

Slide 3
The Workers: In a previous online study, 400 U.S. adults 200 men and 200 women (referred to as

workers) participated in a **math and science test** consisting of 20 multiple-choice questions.
The test assessed their skills in areas such as arithmetic reasoning, mechanical comprehension, math knowledge, general science, and assembling objects.
The 400 workers were **demographically representative of the U.S. adult population**.
Your Task: You are asked to estimate how some of these workers performed on the math and science test compared to other workers.
The more accurate your predictions, the better your chances of earning the bonus.

Slide 4
To make your predictions, you will be provided with basic information about each worker in the form of a short CV. This will include their gender, education level, and month of birth.



Worker CV	
Gender	Female/Male
Education Level	No Bachelor's degree/ Bachelor's degree or higher
Month of Birth	Even/Odd
Citizenship	U.S.

- **Gender**: All participants in the sample identified as either male or female. There were an equal number of male and female workers.
- **Education Level**: Approx. half of the workers held at least a Bachelors degree, while the other half had a lower degree or no degree.
- **Month of Birth**: Approx. half of the workers were born in an odd-numbered month (January, March, May, July, September, November), while the other half were born in an even-numbered month (February, April, June, August, October, December).
- **Citizenship**: All workers were from the U.S.

Slide 5
Which of the following statements is true?

- The sample of "workers" in the previous online survey includes more men than women.
- The sample of "workers" in the previous online survey includes more women than men.
- The sample of ’workers’ in the previous online survey is demographically representative of the U.S. population, with a balanced distribution in terms of gender (female/male), education level (at least a Bachelors degree/no Bachelors degree), and month of birth (even/odd).

Slide 5
Which of the following statements is true?

- The test covered topics in arithmetic reasoning, mechanical comprehension, math knowledge, general science, and assembling objects.
- The test assessed verbal reasoning skills.
- The test measured second language proficiency.

Slide 7
How can you maximize your bonus payment?

- By giving random answers.
- By being as accurate as possible in my predictions.
- There is no chance of a bonus payment.

Slide 8

Part 1: On the following screens, you will see the CVs of workers. You do not know their performance. Your goal is to estimate the likelihood that their performance is in the **top half ("Top Performer")** or the **bottom half ("Low Performer")** relative to the performance of all other workers in the 400-worker sample.

In other words, you are asked to estimate how likely it is that this worker's performance will be in the top 50%.

The more accurate you are, the more likely you are to receive an additional $5. To learn more about the exact bonus payout rule and how it ensures that it is best for you to report your most accurate guess, click here.

How to maximize your chances of earning the bonus?

- If you think the worker is likely to be a Top Performer, move the slider to the right toward 100. The more certain you are, the farther to the right you move it.
- If you think the worker is likely to be a Low Performer, move the slider to the right toward 0. The more certain you are, the farther to the left you move it.
- In any other case, keeping the slider in the middle will increase your chances of earning a bonus.

[Likert Scale: 0%100%]

Slide 9

Payment rule: If your prediction exceeds a randomly drawn number X between 0 and 100, you get paid $5 if the worker you assessed is a Top Performer and $0 otherwise.

If your prediction is below or equal to the randomly drawn number X, then you get $5 with probability $X\%$ and $0 otherwise.

By stating your most accurate guess, you can guarantee that no matter what the draw ends up being, you will be paid according to the option that has the higher probability of paying you an additional $5: the bet on your score or the lottery.

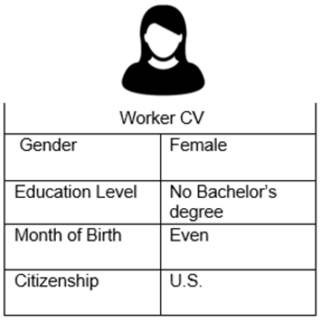
Slide 10: Comprehension Question

What is a Top Performer?

- The worker with the highest score on a test.
- A worker whose performance score ranks in the top 50% among all 400 workers.
- A worker who answered at least 75% of the questions correctly.

Slides 10 through 17: (Worker evaluation, random order)

Example of female worker:

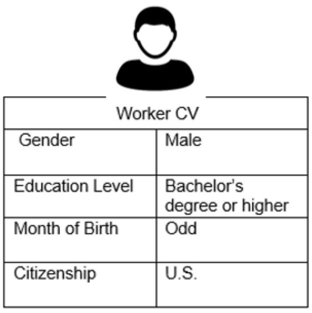


Worker CV	
Gender	Female
Education Level	No Bachelor's degree
Month of Birth	Even
Citizenship	U.S.

How likely is it that she is a Top Performer? *Likert scale from 0% (very unlikely) to 100% (very likely).*

Slides 10 through 17: (Worker evaluation, random order)

Example of female worker:



Worker CV	
Gender	Male
Education Level	Bachelor's degree or higher
Month of Birth	Odd
Citizenship	U.S.

How likely is it that he is a Top Performer? *Likert scale from 0% (very unlikely) to 100% (very likely).*

....

Slide 18

You will now see predictions generated by a machine learning algorithm about whether these workers are Top or Low Performers.

The algorithm was developed using data and performance results from the remaining 392 workers in the sample of 400, and then applied to predict the performance of the eight workers you just evaluated.

Slide 19 [BASELINE CONDITION]: Features the Algorithm Uses to Make Predictions

The following features were used by the *baseline* machine learning algorithm to predict worker performance:

- **Education level** (Bachelors degree or higher / no Bachelors degree)
- **Month of birth** (even / odd)
- **Gender** (female / male)

Performance on a similar test: Before taking the math and science test, workers took another 20-question test that also covered topics such as arithmetic reasoning, mechanical comprehension, math knowledge, general science, and assembling objects. Workers who were Top Performers on this pre-test also tended to be Top Performers on the math and science test, while those who were Low Performers on this pre-test typically remained Low Performers on the math and science test. The algorithm uses performance on this pre-test to predict performance on the math and science test.

You now have the opportunity to revise your previous answers. **Only the answers you provide on the following screens will be considered for the $5 bonus payment, not the ones you previously submitted.**

[Treatment Group: Gender]

Gender is not included as bullet point. Below the bullets, it stated: *"Gender is not included."*

[Treatment Group: Month of Birth]

Month of Birth is not included as bullet point. Below the bullets, it stated: *"Month of birth is not included."*

Slide 20 [BASELINE CONDITION]:

This is how the algorithm works:

$$\operatorname{logit}(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Note: In the experiment, participants saw this equation with arrows and labels that explained each component in simple terms.

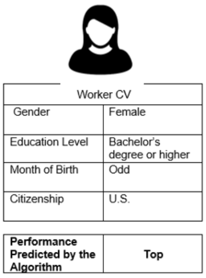
[Treatment Group: Gender]

Gender is not part of the model. Below the equation, it stated: *"Gender is not included."*

[**Treatment Group: Month of Birth**]
Month of birth is not part of the model. Below the equation, it stated: *"Month of birth is not included."*
Slide 21
Which of the following variables does the algorithm use to predict whether the worker is a Top or Low Performer?

- Age
- Gender
- Height
- Performance on a similar test
- Month of birth (even/odd)
- Marital status
- Education level
- SAT score

Slides 22-29 (repeat 10-17 incl. performance prediction, same order as 10-17):
Would you like to change your previous answer?



Worker CV

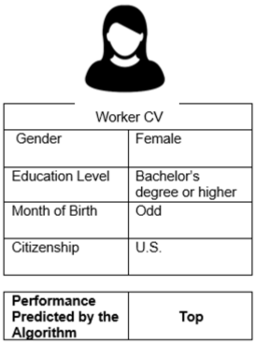
Gender	Female
Education Level	Bachelor's degree or higher
Month of Birth	Odd
Citizenship	U.S.

Performance Predicted by the Algorithm	**Top**

Note: This slide shows one example of the eight prediction screens presented to participants.
Slide 30: Instructions for Part 2
You have completed Part 1.
Part 2 will take about 3 minutes to complete. You will first be asked if you would like to hire the workers you just evaluated to work with you (you will make a total of 8 choices). You will then be asked to complete a simple and short work task. **Bonus payment in Part 2:**

- If you do not solve the work task correctly, you will receive $0.
- If you solve the task correctly, one of your hiring choices will be randomly selected and your payment will depend on whether you hired the worker to work with you:
  - If you do not hire the worker, you get $2.50.
  - If you hired the worker, and they are a Top Performer, you will get $5, and they will get $1.
  - If you hired the worker, and they are a Low Performer, you will get $0, and they will get $1.

If you hire the randomly selected worker, they will receive the fixed payment regardless of their performance.
Slides 22 through 29: Hiring decisions for the same eight workers (same order as in part 1)



Worker CV	
Gender	Female
Education Level	Bachelor's degree or higher
Month of Birth	Odd
Citizenship	U.S.

Performance Predicted by the Algorithm	**Top**

Would you like to hire the worker?

- Yes
- No

If this hiring choice is randomly selected:

- If you solve the simple task correctly and hire her, you get $5 if she is a top performer and $0 if she is a low performer.
- If you solve the simple task correctly and do not hire her, you get $1.
- If you hire her, she will get $1 in any case.

Slide 30

Your task is to solve a short question. You have **45 seconds** to answer it.

Question:

A cookie and a peppermint cost $1.10 in total. The cookie costs a dollar more than the peppermint.

How much does the peppermint cost (in cents)?

(Please omit writing “cents” and only write in the corresponding number (e.g., 0,1,2,).)

Answer: _______________ cents

Slide 31

You’ve completed part 2.

For the following question you can receive a bonus payment of **$1**, if you make an accurate prediction.

Slide 32

Imagine applying the algorithm to 100 workers. On average, how many predictions would be correct? [Insert number 0-100 here].

To increase your chances of receiving an extra $1, make your predictions as accurate as possible.

To learn more about the exact bonus payout rule that ensures accuracy leads to higher bonus payouts, click here.

Slide 33

Please take a moment to answer a few basic questions before finding out how many questions you answered correctly and the amount of your bonus payment.

Slide 34

How much do you agree with the following statements? Participants were asked to indicate their level of agreement using a 6-point scale:

> *Likert scale here: Strongly agree, Rather agree, Neutral, Rather disagree, Strongly disagree, I don't know (slides 34-36)*

- There should be caution in replacing important people tasks with technology because new technology is *not dependable.*
- Whenever something gets automated, you need to check carefully that the machine or computer is not making mistakes.
- When searching for a job online, job offers displayed may vary from person to person despite the same search entry.
- I can assess what the limitations and opportunities of algorithms are.

Slide 35

- The use of algorithms that classify people based on certain criteria can lead to systematic discrimination of some people.
- The exclusion of variables from the algorithm's input data will result in less accurate predictions.
- The exclusion of variables from the algorithm's input data is to ensure that the algorithm is not discriminatory.

Slide 36

- Discrimination against women is no longer a problem in the U.S.
- Women are getting too demanding in their push for equality
- Over the past few years, women have gotten more economically than they deserve

Slide 37

Demographic questions

Slide 38

Thank you for participating in this study.

The predictions were entirely accurate for the eight workers you evaluated.
On average, the algorithm correctly predicts the performance of 74 out of 100 randomly selected workers.

- There were no significant differences in average performance on the math and science test based on gender or birth month.
- Participants with at least a Bachelor's degree typically scored higher than those without one.
- The algorithm based its predictions only on performance on the other math and science test, which was the best predictor of results.
- The correct answer to the question you solved is: **5**.

Your bonus payment is __________.

Your total payment is __________.

Chapter 4.

Social Disparities in Digital Skills: Evidence from Germany

Abstract

This paper documents gender and socioeconomic gaps in digital skills relevant to the labor market, using a representative German household sample. Men and individuals with a higher level of education demonstrate greater proficiency. Both groups also hold more optimistic beliefs about outperforming others, conditional on actual skills. These belief gaps are not driven by overconfidence, but by underconfidence among women and individuals with lower education backgrounds in the upper tail of the skill distribution. Early-life socioeconomic background is not significantly associated with adult digital skills or beliefs.

Keywords: Digital Skills; Technological Transformation; Digital Inequality

JEL Classification: D63; I24; J24; O33

4.A Introduction

The rapid adoption of new technologies across firms and industries is fundamentally shifting the skills required in the labor market (Acemoglu, 2002; Autor et al., 2003; Bick et al., 2024). Demand for digital skills, in particular, is rapidly increasing as workers are now expected to interact with, monitor, and adapt to new technologies (Bick et al., 2024; Carvajal et al., 2024). These skills are already associated with rising labor market returns and are predicted to become a determinant of inequality (Alekseeva et al., 2021; Jackman et al., 2021; Acemoglu et al., 2022; Noy and Zhang, 2023; Eloundou et al., 2024; Gathmann et al., 2024; Brynjolfsson et al., 2025). Despite their growing importance, little is known about their distribution in the population. This is partly because such skills are typically acquired outside formal education systems and redefine themselves with technological progress. It thus remains unclear whether this shift will widen or narrow existing inequalities, and how it interacts with established mechanisms of inequality, such as intergenerational transmission.

This chapter presents new evidence on the distribution of digital skills in the German working population. It uses exclusive data from a survey module implemented in the German Socio-Economic Panel Innovation Sample (SOEP-IS)[1], which enables the study of both cross-sectional disparities across groups and longitudinal links to early-life SES. The module elicits individuals (i) *digital skill levels*, and (ii) two distinct sets of *beliefs*: perceived skill advantage relative to others (of their age group and of the general population), and expectations about how technological change will affect ones own labor market prospects and other life domains. The

[1] This module was developed and submitted in collaboration with Fabian Kosse and Tim Leffler.

paper focuses on disparities by gender and socioeconomic status (SES), both current and early-life.

Digital skills are measured using items adapted from the *Youth Digital Skills Indicator (yDSI)*[2], which is a cross-validated instrument designed to measure several digital skills dimensions in large, representative populations (Helsper et al., 2020; ySKILLS EU Project, 2024). This study focuses on yDSI items most relevant to job performance, specifically those measuring programming, information navigation, and technical and operational skills. Beliefs about own abilities relative to others are central to economic decision-making (Barber and Odean, 2001; Buser et al., 2014; Bordalo et al., 2019; Owen, 2023; Exley and Nielsen, 2024; Roussille, 2024). In the labor market, such beliefs influence whether individuals apply for jobs, choose to compete, negotiate, invest in skill development, or self-promote. Holding abilities constant, women and individuals from lower socioeconomic backgrounds tend to be less confident, in line with stereotypes associated with the domain (Bénabou and Tirole, 2002; Coffman et al., 2021a; Exley and Kessler, 2022; Coffman et al., 2024b; Roussille, 2024; Coffman et al., 2024a). Despite their relevance, these belief gaps have received limited attention in the context of the skill shift on the labor market due to technological change. Yet in precisely this setting, where requirements are dynamic and objective assessments of ability are limited, beliefs about ones own and others skills may be even more consequential (Kaniel et al., 2010). Beliefs about the expected impact of technology on ones future labor market prospects and other life domains are elicited because optimism about one's own future standing has been shown to positively influence forward-looking economic decisions, such as investment, educational choices, and entering into competition (Brunnermeier and Parker, 2005; Puri and Robinson, 2007; Wiswall and Zafar, 2015; Mueller et al., 2021). For example, people with higher income or higher education are more optimistic about future macroeconomic developments (Das et al., 2020), significantly impacting dispariteis in economic outcomes over the long-run.

The data show significant gender and socioeconomic disparities in both digital skills and related beliefs. Men and individuals with higher education and income levels exhibit significantly higher digital skill levels. The gender gap is particularly pronounced in advanced areas, such as programming. The most striking finding pertains to disparities in beliefs: men and higher-educated individuals are significantly more likely to believe they outperform others, conditional on actual skills. This applies for both reference groups, comparisons with peers from the same age group and individuals from the general German population. This gap in beliefs does not appear to be driven by overconfidence, but rather by underconfidence among

[2] The youth Digital Skills Indicator (yDSI) was initiated by the European Commission and is an extensively validated survey instrument designed to measure digital skills among young adults across Europe, assessing competencies in areas such as data literacy, privacy, and content creation.

women and lower-educated individuals at the upper tail of the skills distribution. Men and higher-educated individuals with high skills hold better calibrated beliefs.

As for optimism about how the technological change will impact one's own labor market prospects, those with higher education levels and higher income are significantly more optimistic. However, once digital skills are controlled for, these associations largely disappear, and skills become the only significant predictor. This suggests that skills, rather than SES background, primarily drive personal optimism about future labor market prospects amid technological change. Finally, early-life SES, as measured by parental education, is not significantly associated with digital skills or these beliefs in adulthood.

The findings on skill disparities are consistent with prior evidence on the unequal adoption of technology across gender and socioeconomic groups (Venkatesh et al., 2000; Benyishay et al., 2020; Bertrand, 2020; Breda et al., 2023). Specifically for generative AI, Carvajal et al. (2024) document a significant gender gap in usage among students in Norway, with female students, especially top performers, opting out. This behavior appears to be driven by honesty and compliance concerns: the gap closes when AI use is explicitly permitted and widens under bans. Among workers, Humlum and Vestergaard (2025) show in a Danish sample that women are significantly less likely (16 pp.) to use generative AI at work, and that usage increases with income, indicating parallel socioeconomic disparities in adoption. Bick et al. (2024) report similar findings for the US context, where adoption of generative AI is higher among more educated, younger, and male workers, though the gender gap is smaller (around 9 pp.). The findings of this study also complement the extensive literature on biased technology-driven job displacement that exacerbates existing inequalities (Acemoglu and Autor, 2011; Frey and Osborne, 2017; Acemoglu and Restrepo, 2020; Aksoy et al., 2021; Autor et al., 2024), by pointing to disparities in tech-complementary skills as a parallel mechanism. Cazzaniga et al. (2025) present recent global data showing that female workers with a low-SES background are at disproportionately high risk of job loss, as they are more likely to work in AI-exposed occupations.

Disparities in individuals' beliefs about their own skill levels, relative to others, align with prior research showing beliefs respond to social stereotypes (Coffman, 2014; Bordalo et al., 2019; Coffman et al., 2024a; Campos-Mercade and Mengel, 2024). The findings here can be interpreted through the model in Bordalo et al. (2016), which incorporates the kernel of truth property. This property posits that while stereotypes are rooted in actual average differences between groups, these differences are exaggerated in perceived beliefs. Specifically, the skills of individuals from groups performing better on average are overestimated, while the skills of individuals from groups performing worse on average are underestimated. In short, the "kernel of truth" predicts stereotypes will exaggerate perceived performance gaps between groups. When applied to this study's context, the data shows that, on average, men and individuals with higher education levels do indeed show higher proficiency levels on average. Thus, the direction of the belief bias aligns with the true group-level variation. However, the perceived gap is exaggerated: when actual skills are held constant, women and individuals with lower levels of education tend to underestimate their skills relative to others. Contrary to existing literature (Bordalo et al., 2019; Coffman et al., 2024a), the bias here does not reflect overconfidence among higher-skilled groups, but rather underconfidence among individuals in groups that show lower skills on average (in line with the mechanism in Campos-Mercade and Mengel (2024)).

The finding that optimism about the labor market implications of technological change is associated with individual skills rather than socioeconomic background contrasts with earlier research showing a strong relationship between socioeconomic conditions and beliefs about future economic outcomes and upward mobility (Kearney and Levine, 2014; Genicot and Ray, 2017). Instead, it is more consistent with recent evidence suggesting that expectations are determined primarily by individual characteristicssuch as skills, knowledge, and endowmentsrather than group membership (DAcunto et al., 2019; Dacunto et al., 2023). These results indicate that beliefs about how technological change will affect ones future are less closely linked to persistent socioeconomic circumstances and may be more responsive to individual learning and skill acquisition.

Finally, the lack of a significant association between early-life socioeconomic status and either digital skills or related beliefs suggests that intergenerational inequality plays a limited role in shaping individuals ability to adapt to a digitalized labor market. This finding contrasts with a large literature documenting strong and persistent links between childhood socioeconomic environments and the formation of human capital, including cognitive and non-cognitive skills as well as long-term economic outcomes (Heckman, 1976; Becker and Tomes, 1979; Erikson and Goldthorpe, 2002; Heckman, 2011). One explanation lies in the nature of digital skills: unlike traditional academic skills, they are frequently acquired outside formal education systems and change in nature as technology advances. Also, for much of the sample, digital tools were not present during early life and diffused across household backgrounds relatively evenly once adopted (Keller, 2004).

Methodologically, this paper adds to a growing literature in empirical economics that leverages large-scale, representative survey data to examine heterogeneity in preferences, beliefs, and economic behavior across socioeconomic groups (Adda et al., 2022; Andre et al., 2022). The German Socio-Economic Panel (SOEP), used in this study, is one of the most commonly applied and extensively validated longitudinal dataset (Fischbacher et al., 2024). It is increasingly employed to measure beliefs and preferences in representative populations (see Stantcheva (2023) for a review and guide), including those related to the labor market, e.g., pertaining to perceived outside options (Jäger et al., 2024) and reservation wages (Mui and Schoefer, 2025). In this study, beliefs are elicited without monetary incentives. While this approach departs from standard economic literature, it is increasingly accepted when using large-scale population panels where incentives may be prohibitively costly, infeasible, or introduce selection effects.

The remainder of this chapter is organized as follows. The next section describes the measures and samples. Section 3 presents results. Section 4 discusses the findings and Section 5 concludes.

4.B Measures and Sample

This section describes the measures and samples employed in the analysis. It begins with the survey instrument used to measure digital skills and beliefs, followed by a description of the representative German household sample drawn from the German Socio-Economic Panel (SOEP).

4.B.1 Measuring skills and beliefs: Survey items

Digital skills are measured using survey items from the "Youth Digital Skill Indicator" (yDSI), a survey instrument developed to assess digital proficiency among young adults across Europe, initiated by the European Commission (Helsper et al., 2020; van Deursen et al., 2023; LSE Media and Communications, 2024). The validity of these stated skill measures is supported by cognitive interviews and performance-based validation, in which participants completed digital tasks under controlled conditions to benchmark self-reports against actual performance. The yDSI scale was specifically designed for use in large-scale, representative population surveys and includes only items with high construct validity.[3]

The items included in this study are drawn from three of the five digital skill dimensions in the yDSI, selected for their relevance to labor market applications.[4] All questions from the selected dimensions are used in their original German form (Appendix Table A21), with English translations provided in Appendix Table A22. Each item is measured on a 5-point Likert scale, with higher values indicating greater proficiency.[5] The three selected skill dimensions are:

- *Technical and Operational Skills* (6 survey items; e.g., “I know how to adjust privacy settings.”).

[3] The original German phrasing of the survey items follows Waechter et al. (2021).

[4] Due to space constraints in the SOEP questionnaires, the full yDSI battery (24 items across five dimensions) was not implemented.

[5] Responses marked as no answer are coded as missing (NA). The response I dont know is coded as 1, reflecting low proficiency.

- *Programming Skills* (1 survey item: “I know how to use programming languages such as XML, Python, Java, C++.”).
- *Information Navigation and Processing Skills* (6 survey items; e.g., “I know how to choose the best keywords for online searches.”).

Beliefs about relative digital skill advantage are measured by asking respondents to estimate their percentile rank based on two reference groups: (i) their age cohort and (ii) the general population. Respondents are asked to estimate the proportion of each reference group with weaker digital skills than those they possess, only based on the dimensions assessed in the preceding items. The survey questions are phrased as follows: When we talk about digital skills, we mean the dimensions you just rated. What do you estimate: (i) What percentage of your age group in Germany currently has less developed digital skills than you? [0100%], and (ii) What percentage of the general population in Germany currently has less developed digital skills than you? [0100%].[6]

The two reference groups, age group and general population, serve distinct purposes. Comparisons with the age cohort reflect one’s perceived standing relative to direct labor market competitors. In contrast, comparisons with the general population provide a broader benchmark and allow for cross-generational comparisons that also account for age-related differences in digital skill levels. The general population reference group also allows for estimating the accuracy of beliefs about others’ skills based on the representative SOEP data.

The survey concludes with four items eliciting subjective beliefs about the implications of digitalization across distinct life domains. Respondents are first presented with the prompt: Thinking about the progressive digitalization in various areas of life, what do you think: in which areas do the risks outweigh the opportunities, and in which areas do the opportunities outweigh the risks? The assessed domains include: (i) opportunities in the labor market; (ii) relationships with friends and family; (iii) organization of free time; and (iv) the overall development of society. Each item is rated on a 5-point Likert scale, ranging from 1 (digitalization is a risk) to 5 (digitalization is an opportunity).[7]

[6] The original German wording is: "Wenn wir über digitale Fähigkeiten reden, meinen wir die Dimensionen, zu denen Sie sich gerade selbst bewertet haben. Was schätzen Sie: (i) Wie hoch ist zurzeit der Anteil in Ihrer Altersgruppe in Deutschland, der schlechter ausgeprägte digitale Fähigkeiten besitzt als Sie?" and "(ii) Wie hoch ist zurzeit der Anteil der allgemeinen Bevölkerung in Deutschland, der schlechter ausgeprägte digitale Fähigkeiten besitzt als Sie?"

[7] Original German wording: "Wenn Sie an die fortschreitende Digitalisierung in den verschiedenen Bereichen des Lebens denken, was meinen Sie, in welchen Bereichen überwiegen die Risiken und in welchen Bereichen überwiegen die Chancen der Digitalisierung? (i) Ihre Möglichkeiten auf dem Arbeitsmarkt; (ii) Ihre Beziehung zu Freunden und Familie; (iii) Die Gestaltung Ihrer Freizeit; (iv) Die Entwicklung der Gesellschaft."

4.B.2 German Socio-Economic Panel (SOEP)

The digital skills measures were implemented as a new, one-time survey module within the Innovation Sample of the German Socio-Economic Panel (SOEP-IS) (e.g., Richter and Schupp, 2012).[8] The module was fielded in the 2023 wave of the SOEP-IS, conducted between April and July 2023. A key advantage of the SOEP-IS is its linkage to the core SOEP panel, which provides both cross-sectional socioeconomic (SES) measures and longitudinal background data, including parental characteristics that proxy early-life SES.

The SOEP subsample assigned to this module includes 1,001 respondents, of whom 992 completed the digital skills items. The sample is 49.4% male, 50.2% female, and 0.4% not identified, with a mean age of 59 years and an average gross monthly income of approximately EUR 3,500. Appendix Tables A4–A7 provide additional information on the distributions of age, income, and education. As the analysis focuses on labor market inequality, the sample is restricted to non-retired individuals in all subsequent analyses.

4.C Results

This section presents findings on gender and socioeconomic disparities in digital skill levels (Section 3.1) and beliefs about digital skill levels (Section 3.2). Section 3.3 presents findings regarding beliefs about the implications of digitalization for one's labor market prospects and other life domains.

4.C.1 Skills: Gender and socioeconomic gaps

Table A1 reports OLS estimates with standardized digital skills as the dependent variable and gender and socioeconomic characteristics as regressors.[9] Digital skills (std.) are constructed as the unweighted average of thirteen items, each of which is standardized across all respondents. The sample excludes retired individuals to focus on the active labor force.[10]

[8] The SOEP Innovation Sample (SOEP-IS) is a sub-sample of the SOEP used to field new, specialized survey modules. While the SOEP-IS sample is longitudinal, individual modules are typically implemented once, but can be linked to longitudinal data from the core SOEP. New modules are selected annually through a competitive application process at the German Institute for Economic Research (DIW). The application for this module was submitted in 2022 and selected for implementation in 2023.

[9] Standard errors are clustered at the household level. In this sample, 43.4% of respondents share a household with at least one other participant, which can introduce within-household correlation due to shared environments or response behavior.

[10] This restriction accounts for the largest drop in observations. Roughly one-third of adults in Germany are retired (21 million of 70 million; Statistisches Bundesamt 2023).

Table A1: Disparities in Digital Skills (Std.)

	Dependent Variable: Digital Skills (std.)		
	(1)	(2)	(3)
Male	0.215***	0.222***	0.164***
	(0.050)	(0.051)	(0.055)
Education level	0.072***	0.069***	0.050***
	(0.009)	(0.009)	(0.011)
Parents' education level		0.010	0.003
		(0.012)	(0.013)
Income			0.081***
			(0.027)
Observations	431	375	347
Controls	Age	Age	Age
R^2	0.250	0.242	0.210
Adj. R^2	0.244	0.235	0.199

Note: OLS regressions with standard errors clustered at the household level (shown in parentheses). The dependent variable is standardized digital skills, calculated as the unweighted average of all thirteen survey items on digital skills, each standardized across respondents. *Male* is a binary indicator equal to 1 for male respondents. *Education level* refers to years of formal education. *Parents' education Level* refers to the years of education of the more highly educated parent. *Income (std. within ten-year age group)* denotes gross monthly income, standardized within ten-year age cohorts to account for income variation over the life cycle. The sample excludes individuals who are retired. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Model 1 estimates the association of gender and education with digital skills. These variables are introduced first, as both are exogenous to skill acquisition and widely studied predictors of digital inequality. Education is measured in total years of formal formal educationfor example, approximately 18 years for a university degree (i.e., *Diplom* or *Masters*) and about 9 years for the lowest secondary school qualification (i.e., *Hauptschulabschluss*). The results indicate that men score significantly higher than women ($\beta = 0.215$, $p < 0.001$), that each additional year of education is associated with a 0.072-point increase in the composite digital skills variable ($p < 0.001$). Together, gender and education explain approximately 25% of the variation in standardized skill levels. These variables are thus key predictors of skill differences.

Model 2 adds parental education to test whether early-life socioeconomic background predicts digital skills in adulthood. Parental education is measured as the total years of formal education completed by the more educated parent.[11] The coefficient on parental education is small and statistically insignificant ($\beta = 0.010$, $p = 0.43$). The estimates for gender ($\beta = 0.222$, $p < 0.001$) and education ($\beta = 0.069$, $p < 0.001$) remain stable in magnitude and significance. These results suggest that digital skills in adulthood are not strongly linked to parental education, i.e.,the early-life SES background, and that intergenerational inequality plays a limited role in this domain.

[11] Consistent with prior work, the more educated parent is used rather than the average or lower value, as this parent tends to exert greater influence on household resources and childrens educational outcomes (Erikson and Goldthorpe, 2002; Chetty et al., 2014).

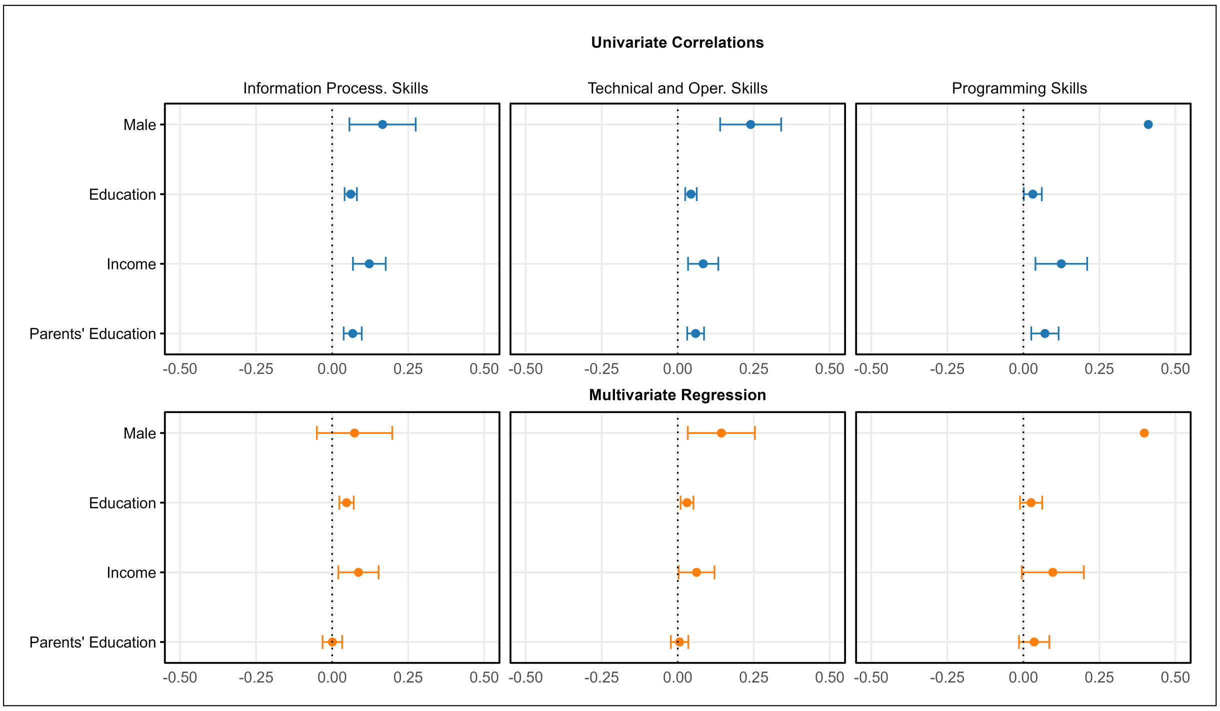
Model 3 adds income as a measure of current socioeconomic status to examine its association with digital skills.[12] Income is standardized within ten-year age cohorts to account for life-cycle variation and to reduce collinearity with age. Unlike gender and education, income may be endogenous to digital skills, as higher digital proficiency can influence labor market outcomes and earnings. The results show a positive and statistically significant association between income and digital skills ($\beta = 0.081$, $p = 0.003$).[13] This indicates that individuals with higher earnings tend to exhibit higher proficiency. The coefficients on gender ($\beta = 0.164$, $p = 0.003$) and education ($\beta = 0.050$, $p < 0.001$) remain statistically significant and large, which suggests that the disparities associated with gender and education persist even after conditioning on income. The lack of significance for parental education ($\beta = 0.003$, $p = 0.77$) remains stable.

Taken together, the results imply that gender and education are strong predictors of digital skills, while income is positively associated with skill levels but may be subject to reverse causality. Parental education, by contrast, shows no significant association. These findings suggest that individual-level characteristics in adulthood account for more variation in digital proficiency than early-life socioeconomic background. All results are robust to alternative model specifications, including the use of a non-standardized dependent variable, a binary indicator for *Abitur* (university entrance qualification), and two alternative measures of income, as shown in Appendix Tables A8–A10. It should be noted that when education is conceptualized as a binary indicator for *Abitur* in Appendix Table A9,the coefficient increases and exceeds the magnitude of the gender coefficient.

[12] Variance inflation factors (VIFs) for all covariates in Model 3 are low: 1.12 for gender, 1.30 for education, 1.15 for parental education, 1.30 for income, and 1.04 for age. As all values fall well below conventional thresholds (typically VIF $>$ 5), multicollinearity is not a concern in this specification. The relatively low VIFs may reflect the standardization of income within age cohorts and the use of years of formal education as a continuous measure.

[13] The lower R^2 in Model 3 may is based on the smaller estimation sample due to missing income data and does not imply a decline in explanatory power of income.

Figure A1: Disparities by Skills Dimension



Notes: The figure presents results for three dimensions of standardized digital skills: *Information Navigation and Processing*, *Technical and Operational Skills*, and *Programming Skills*. The top row shows univariate correlations; the bottom row shows coefficients from multivariate regressions, controlling for age and including household fixed effects. Each skill dimension is standardized across respondents. Information and Technical Skills are based on the average of six standardized self-assessment items each. Programming Skills are measured using a single item, standardized across respondents. *Male* is a binary indicator equal to 1 if the respondent is male. *Education* refers to completed years of formal education. *Income* is standardized within ten-year age groups. *Parents' Education* denotes the years of education completed by the more highly educated parent.

Yet these results do not capture potential heterogeneity across the three distinct dimensions of digital skills: (i) *Information Navigation and Processing*, (ii) *Technical and Operational Skills*, and (iii) *Programming Skills*. These domains differ in nature and complexity, and may therefore vary in their relevance for labor market outcomes (Acemoglu et al., 2022; Kalyani et al., 2025). Figure A1 presents univariate correlations and multivariate regression estimates separately for each dimension. In the univariate models, all predictors are positively and significantly associated with skill levels. The gender variable *Male* exhibits the greatest variation in predictive strength, with particularly strong associations in the more advanced domains of *Technical* and *Programming* skills. *Education* and *Income* are also positively associated with all three dimensions, though their coefficients are more consistent in size across these dimensions. The smaller size of the coefficients may be due to the fact that both variables are measured continuously. Appendix Tables A11 and A12 provide summary statistics for each of the thirteen elicited digtial skills items for the full sample and for the non-retired (working-age) population, respectively.

The next section investigates whether disparities in actual skills also extend to individuals belief on their *relative advantage* in outperforming others.

4.C.2 Beliefs about skills: Gender and socioeconomic gaps

Table A2 reports OLS estimates of individuals beliefs about their percentile rank in the digital skills distribution, measured relative to the general population (Models 1 and 2) and to others in the same age group (Models 3 and 4) in Germany. Models 1 and 3 include gender, education, and income as predictors; Models 2 and 4 additionally control for actual skill levels. In both reference groups, men report significantly higher perceived ranks than women (general population: $\beta = 0.341$, $p < 0.001$; age group: $\beta = 0.444$, $p < 0.001$). The gender gap is larger when individuals compare themselves to individuals from their age group (an increase of 0.103 in the coefficient). Similarly, higher education levels are positively associated with beliefs in both reference groups (general population: $\beta = 0.082$, $p < 0.001$; age group: $\beta = 0.075$, $p < 0.001$). In contrast, parental education and income show no significant links with perceived skills advantages in either reference group.

Table A2: Disparities in Beliefs about Skills Advantage

	DV: Belief About Relative Digital Skill Advantage (Std.)			
	(1) Population	(2) Population	(3) Age Group	(4) Age Group
Male	0.341***	0.232***	0.444***	0.352***
	(0.090)	(0.088)	(0.092)	(0.087)
Education level	0.082***	0.051***	0.075***	0.050**
	(0.019)	(0.018)	(0.019)	(0.019)
Parents' education level	0.029	0.030	0.007	0.008
	(0.022)	(0.021)	(0.024)	(0.024)
Income	0.020	-0.020	0.062	0.023
	(0.053)	(0.050)	(0.062)	(0.062)
Digital skills		0.615***		0.524***
		(0.085)		(0.095)
Observations	344	344	348	348
Controls	Age	Age	Age	Age
R^2	0.258	0.365	0.141	0.216
Adj. R^2	0.247	0.354	0.129	0.202

Notes: OLS estimates with standard errors clustered at the household level (in parentheses). Dependent variables of Models (1) and (2) are standardized measures of subjective skill ranking relative to the general population. Dependent variables of Models (3) and (4) are standardized measures of subjective skill ranking relative to others in the same age group. Survey question: : What percentage of the German general population [those in your age group in Germany] do you believe has weaker digital skills than you? *Male* is a binary indicator equal to 1 for male respondents. *Education Level* refers to years of formal education. *Parents' Education Level* refers to the years of education of the more highly educated parent. *Income (std. within ten-year age group)* denotes gross monthly income, standardized within ten-year age cohorts to account for income variation over the life cycle. All models control for age. The sample excludes retired individuals. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Given the similarity between these results and the previous ones on actual skill levels, one might expect that belief differences simply reflect underlying skill differences, i.e., that individuals accurately estimate and report their advantage. However, when controlling for actual digital skills in Models 2 and 4, the effects of gender and education remain large and statistically significant (general population: $\beta = 0.232$, $p < 0.001$; age group: $\beta = 0.352$, $p < 0.001$). Actual skill differences explain approximately 32% of the gender gap in beliefs for the general population comparison and 21% for the comparison with individuals from the same age group. For education, the shares are 38% and 33%, respectively. These results indicate that differences in beliefs about relative skill advantages do not reflect actual skill differences alone, but rather reflect systematic variation across groups in how individuals subjectively evaluate their relative standing. These findings are robust to alternative specifications using different measures of education (Appendix Table A13), and to other measures of income, i.e., income not std. based on age-group but full sample, and (ii) household income (Appendix Table A15).

Are these biases in beliefs driven by overconfidence among men and those with higher education levels or underconfidence among women and individuals with lower education? Figure **??** plots individuals subjective beliefs about their digital skills percentile rank against their actual percentile rank, based on SOEP panel data for the general population.[14] Linear fits are estimated separately by gender (left panel) and education level (right panel) using OLS estimates. Appendix Table A16 and Appendix Table A17 report corresponding regression results.

For gender, as shown in all models in Appendix Table A16, the intercept is negative and highly significant, implying that respondents underestimate their relative skill rank on average. The coefficient on actual skills is also negative ($\beta = -0.62$ to -0.78, $p < 0.001$). This indicates that individuals with higher skill levels are more prone to underestimation, which holds for both genders. However, the extent of underestimation, differs significantly by gender. Women underestimate their skills significantly more than men (Model 2: $\beta = 4.07$, $p = 0.019$), especially at higher skill levels (Model 2 interaction: $\beta = 0.24$, $p < 0.001$; Model 3: $\beta = 0.28$, $p < 0.001$; squared skill interaction: $\beta = 0.005$, $p = 0.079$). For example, at the 90th percentile of actual skill, men rate their relative performance 20.2 percentage points higher than equally skilled women. This difference reflects better calibration, not overconfidence: both genders continue to underestimate their performance, but men do so to a significantly lesser degree. The gender gap in underestimation widens with increasing skill. At lower skill levels, there are no gender differences. At the 40th percentile, the interaction term is not significant ($p = 0.823$, $N = 107$), nor at the 30th percentile ($p = 0.574$, $N = 62$).

[14] Comparisons to age group peers are excluded due to ambiguity in the reference category. The survey item did not specify age thresholds, which prevents a valid interpretation across respondents.

In terms of education, again across all models in Table A17, the intercept is significantly negative, which confirms that individuals, on average, underestimate their relative skill rank. The coefficient on actual skills is again negative and highly significant ($\beta = -0.60$ to -0.65, $p < 0.001$), implying that individuals with higher digital skills tend to exhibit greater underconfidence, consistent with the finding for gaps based on gender groups.

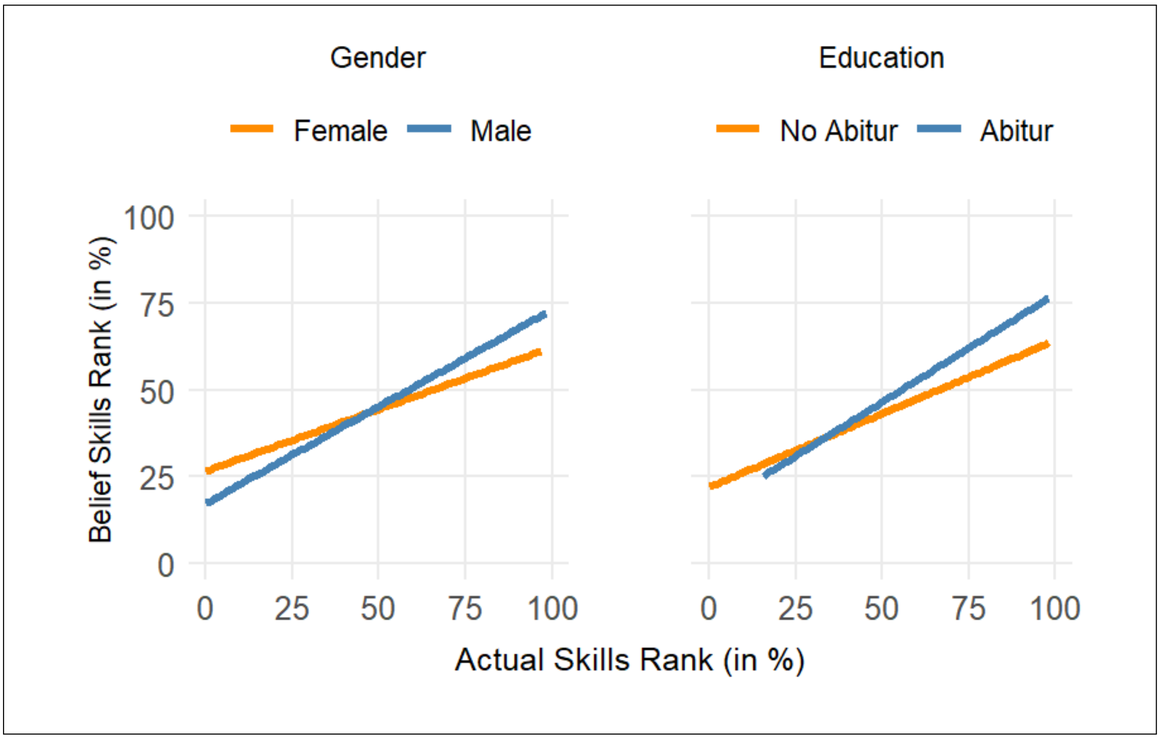


Figure A2: Skills Percentile Rank Relative to the General Population (Actual vs. Belief)

Notes: The figure presents linear trend lines estimated using ordinary least squares. It predicts the relationship between individuals actual digital skill percentiles, calculated from the representative SOEP dataset, to their belief about their skills relative to the population. Belief is measured by the survey question: What percentage of the German general population do you believe has weaker digital skills than you? Both variables represent the percentage of the population performing worse. The left panel (n = 539) shows results by gender. The right panel (n = 444) shows results by education level and distinguishes between those with and without *Abitur* degree, i.e., the highest secondary school qualification in Germany. The sample excludes retired individuals.

However, the extent of this underconfidence varies significantly by education level. In Model 1, respondents with an *Abitur* degree are significantly less underconfident than those without it ($\beta = 6.70$, $p < 0.01$). Model 2 further shows that the interaction between education and actual skills is significant ($\beta = 0.160$, $p < 0.1$). This indicates that the gap is increasing for individuals at the upper skills distribution. Model 3 including quadratic terms confirms this, though the interaction term becomes statistically insignificant.

To illustrate the magnitude of this difference, consider again the 90th percentile of actual skill, i.e., 40 points above the mean. Based on Model 3, the predicted miscalibration at this point is -23.64 percentile points for respondents with *Abitur* and -40.73 for those without, corresponding to a gap of 17.09 points in beliefs despite equal skill levels. As with gender, this does not indicate overconfidence among the highly educated: miscalibration remains negative for both groups when considering the intercept. Rather, the results suggest that individuals with higher education levels, especially those with strong digital skills, have more accurate beliefs of their skill advantage. In contrast, those with lower education levels tend to be significantly more underconfident.

In terms of education, individuals with *Abitur* (the highest secondary education degree in Germany) exhibit significantly less underconfidence than those without, conditional on actual skill level (Model 2: $\beta = 5.92$, $p = 0.002$; Model 3: $\beta = 6.87$, $p = 0.007$). This difference increases with skill; at higher levels, those with *Abitur* underestimate their performance less than those without (Model 2 interaction: $\beta = 0.16$, $p = 0.069$). However, this interaction becomes statistically insignificant in Model 3, which includes a squared term ($p = 0.137$). These results suggest that individuals with higher levels of education report beliefs that more closely align with their actual performance at particularly high skill levels, whereas those with lower education continue to underestimate their skills.

4.C.3 Beliefs about the impact of technology

This subsection examines heterogeneity in beliefs about how they digitalization of the labor market impacts own job market prospects. The results in Models 12 reported in Table A3 indicate that individuals with higher levels of education, and, to a lesser extent, those with higher income are significantly more optimistic. In contrast to the prior results, there is no significant association with gender. Parental education in Model 3 is not significantly linked to those beliefs.

Model 4 includes digital skill levels as an additional explanatory variable. Digital skills are positively and significantly related to optimism ($p = 0.009$). Strikingly, once digital skills are included, the coefficients for education and income decline substantially and lose statistical significance ($p = 0.240$ for education and $p = 0.319$ for income). The adjusted Rš increases from 0.020 in Model 3 to 0.036 in Model 4, which corresponds to an 80 percent improvement in explanatory power.

These results suggest that digital skills can offset the influence of socioeconomic background on beliefs about the impact of technological change. Once skill differences are accounted for, education and income no longer explain variation in optimism. This indicates that differences in the ability to leverage technological change ("preparedness"), rather than background characteristics per se, may shape beliefs about personal future labor market opportunities amid the technological change.

Appendix Tables A18 to A20 present OLS regression results for beliefs about the impact of digitalization on society, personal leisure time, and personal relationships. Digital skills do not significantly predict optimism in these domains, with the exception of a marginally significant positive association for leisure time. There are also no statistically significant differences in these beliefs by gender or socioeconomic background.

Table A3: Beliefs: Impact of Digitalization on Own Labor Market Prospects

	DV: Digitalization is Opportunity for Own Labor Market Success (Belief)			
	(1)	(2)	(3)	(4)
Male	0.018	0.003	-0.004	-0.027
	(0.061)	(0.064)	(0.063)	(0.065)
Education Level	0.035**	0.028**	0.021*	0.014
	(0.012)	(0.014)	(0.012)	(0.012)
Income		0.065*	0.054	0.041
		(0.036)	(0.040)	(0.041)
Parent Education			-0.004	-0.004
			(0.016)	(0.015)
Digital Skills (std.)				0.169**
				(0.064)
Observations	430	375	365	365
Adj. R^2	0.027	0.034	0.020	0.036
Controls	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control

Notes: OLS regressions with standard errors clustered at the household level. The dependent variable is standardized belief that digitalization improves ones labor market prospects. *Male* is a binary indicator equal to 1 for male respondents. *Education Level* is measured in years. *Income* is standardized within ten-year age cohorts. *Digital Skills (std.)* is a standardized measure of actual digital skills. All models control for age and locus of control, measured by agreement with: The outcome of my life is within my control. The sample excludes individuals in retirement. Standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.D Discussion

This paper makes three main empirical contributions to the study of digital inequality, using representative German household panel data. First, it documents gender and socioeconomic differences in digital skills. Individuals with higher levels of education and income, as well as men, are significantly more proficient in all three measured skills dimensions. Digital proficiency is thus disproportionately concentrated among groups with already exisiting labor market advantages. While prior work has emphasized disparities in digital access and usage, this paper shifts the focus to skills, a related but distinct factor that is likely mutually endogenous with technology adoption.

Second, and most strikingly, the paper presents novel evidence that these gender and socioeconomic gaps also apply to beliefs about relative digital skill, i.e., who is confident that they outperform others in when it comes to digital proficiency, even conditional on actual skill endowment. Given prior evidence on the relevance of these beliefs for labor market decisions, such miscalibrated self-assessments about own competitiveness, could imply that equally skilled workers frmo disadvantaged groups may be less likely ot apply for jobs requiring skills, invest in digital skills education, and self-promote htemselves in their relative advantage over others.

Belief distortions of this kind may therefore contribute to the persistence of inequality, over and above differences in measured skill.

Third, leveraging the longitudinal data of two panel datasets, the paper provides new, suggestive evidence that digital skills and related beliefs are not strongly associated with early-life socioeconomic background. While prior research has documented robust intergenerational transmission of human capitalincluding both cognitive and non-cognitive skillsthe results here suggest that such persistence may not extend to the digital domain. Instead, individual-level factors such as current education and income appear to play a larger role in shaping both skill endowments and beliefs. This weaker dependence on parental background may open new avenues for upward mobility in a labor market increasingly demanding digital proficiency.

Disparities in digital skills and confidence in those skills can independently, and especially jointly, exacerbate existing labor market inequalities as technological change quickly reshapes job requirements. These findings underscore the urgent need for targeted policy interventions in several areas: It is crucial to reinforce digital skill development and continuous training from an early stage in the education system. These interventions should account for the needs and expectations of children and students from different gender groups and socioeconomic backgrounds to be effective. Furthermore, it is important to address bias in beliefs about own skills and strengthen the confidence of individuals from disadvantaged groups in their ability to improve their skills through learning. Otherwise, this can hinder their successful labor market participation, upward mobility, and investment in training over the long run. Finally, regarding the demand side, corporate policies should promote on-the-job learning and account for biases in self-assessed skill levels and training needs, especially among individuals from disadvantaged backgrounds. For instance, Exley and Nielsen (2024) show that even when employers are aware of biased self-assessments in own abilities across groups (in their case, men overstating their skills relative to women) they tend to fail to adjust appropriately in hiring and promotion decisions, which ultimately reinforces disparities. This is particularly concerning in light of recent evidence on dramatically rising returns to digital skills, not only in hiring and promotion but also in their investment in worker training (Carvajal et al., 2024; Brynjolfsson et al., 2025).

The findings of this study should be interpreted with caution due to several methodological limitations. First, the analysis relies on unincentivized self-reports of digital skills and related beliefs, which introduces potential measurement error. Although the "Digital Skills Indicator" used in this study was selected for its rigorous validation processes, the data reflects only stated, not revealed, proficiency levels, raising questions about their validity. Additionally, these responses may reflect not only objective skill levels, but also beliefs about relative performance. Even though respondents were not yet asked about comparisons when reporting their own skill levels, social comparisons may already be embedded in these judgments (Bordalo et al., 2019). Nevertheless, these results are informative because, in practice, employers often use self-reported skill measures for hiring and promotion decisions, as objective evaluations are often costly or unavailable.

Second, the measurement of digital skills is constrained by the surveys limited set of skill dimensions, due to space constraints, and some of these may already be outdated. As technology and workplace tools evolve rapidly, skills captured in 2022 and 2023 may have lost relevance, while newer skills, like prompting in generative AI applications, are increasingly valued by employers. Although limited, the items used herecovering data privacy, keyword use, and programming (as a proxy for IT proficiency)-still offer a reasonable proxy of digital proficiency. Third, the sample size is limited because the SOEP subsample used in this study focuses on the active labor force, excluding retirees, and the sample is older on average than the general German adult population. As a result, a larger number of observations than expected had to be excluded from the analysis. Additionally, sensitive sociodemographic variables such as income are frequently missing or may be prone to reporting error and noise.

This study is also limited in that it conceptualizes inequality through the traditional economic dimensions of gender, education, and income. However, other factors, such as age, ethnicity, disability status, geographic location, and migration background, are major sources of discrimination and exclusion from the labor market. A full analysis of these dimensions is beyond the scope of this paper; however, this limitation does not reflect a judgment on their relevance. For example, older individuals have been shown to face discrimination in the labor market (Neumark et al., 2016), and, in parallel, are disadvantaged in the development, applications, and regulatory frameworks of new technologies (Nielsen and Woemmel, 2024).

These considerations point to several directions for future research. First, the robustness of these findings should be assessed using alternative skill and belief measures, cross-country data, and proficiency tests. Second, given the pace of technological change, it is important to continuously update the skill dimensions measured in this study. As shown by Kalyani et al. (2025), the adoption of new technologies is initially associated with high-skilled workers and high-return jobs, but over time draws in lower-skilled workers as the required skill premium declines, e.g., as shown by the recent drop in value of programming skills due to generative AI. Third, building on the long-standing economic literature (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018), future research should identify which digital skills complement emerging technologies and which are likely to become obsolete. Distinguishing between these two is essential for anticipating shifts in labor market demand and designing effective education and training policies to reduce inequality.

4.E Conclusion

The rapid technological change in the workplace may perpetuate inequality, not only through biased automation and job displacement, but also through gaps in who can leverage the new skill demands it creates. Even when technological change is implemented uniformly across sectors and occupations, it may reinforce existing inequalities. Individuals from disadvantaged groups may lack the skills and confidence necessary to compete and invest effectively. Thus, the technology's impact on inequality depends not only on its design and diffusion, but also on who feels capable of benefiting from it. Understanding these dynamics and effectively addressing them with policy responses requires continuous and interdisciplinary research.

4.A Appendix

Table A4: Distribution of Age Groups (SOEP)

Age Category (in years)	*N*	*Share*
25 or younger	28	0.028
2635	87	0.087
3645	116	0.116
4655	124	0.124
5665	212	0.212
6675	233	0.233
76 or older	179	0.179
Missing (NA)	22	0.022
Total	1001	1.000

Notes: The table summarizes the distribution of respondents by age category. *N* is the number of observations; *Share* denotes the fraction of the total sample (N = 1001).

Table A5: Distribution of Individual Gross and Household Net Monthly Income (SOEP)

Income Category	*Individual Gross Income: N*	*Share*	*HH Net Income: N*	*Share*
Below 1,000	73	0.073	37	0.037
1,0001,999	60	0.060	168	0.168
2,0002,999	84	0.084	237	0.237
3,0003,999	120	0.120	189	0.189
4,0004,999	67	0.067	135	0.135
5,0005,999	44	0.044	101	0.101
6,000 or more	57	0.057	103	0.103
Missing (NA)	496	0.496	31	0.031
Total	1001	1.000	1001	1.000

Notes: The table shows the distribution of monthly income (individual and household level) within the SOEP sample (German general population). *N* is the number of observations in each category; *Share* denotes the proportion of the total sample (N = 1001).

Table A6: Distribution of General Education Qualifications

Education Category	*N*	*Share*
Secondary School Certificate (Realschulabschluss)	263	0.263
High School Diploma, Upper secondary education (Abitur)	239	0.239
Lower Secondary School Certificate (Hauptschulabschluss)	165	0.165
Advanced Technical College Entrance Qualification (Fachhochschulreife)	59	0.059
Secondary School (Realschule)	53	0.053
Other Qualification	35	0.035
Left School Without Qualification	15	0.015
Academic Secondary School (Gymnasium)	5	0.005
No Qualification Yet	4	0.004
Technical Secondary School (Fachoberschule)	1	0.001
Missing (NA)	162	0.162
Total	1001	1.000

Notes: The table summarizes the distribution of the secondary education levels within the SOEP sample (German general population). *N* is the number of observations; *Share* denotes the fraction of the total sample (N = 1001).

Table A7: Distribution of Highest Post-Seconddary Education

Education Category	*N*	*Share*
Apprenticeship ("Lehre")	413	0.412
University Degree	179	0.179
Vocational School	73	0.073
Technical School (e.g., Master Craftsman)	66	0.066
Civil Service Training	27	0.027
Other Advanced Degree	9	0.009
No Advanced Degree	57	0.056
Missing (NA)	177	0.177
Total	1001	1.000

Notes: The table summarizes the distribution of the vocational or tertiary educational level within the SOEP sample (German general population). *N* is the number of observations; *Share* denotes the fraction of the total sample (N = 1001).

Table A8: Predictors of Digital Skills (not standardized)

	Dependent Variable: Digital Skills		
	(1)	(2)	(3)
Male	0.310***	0.321***	0.232**
	(0.076)	(0.078)	(0.083)
Education level	0.108***	0.104***	0.075***
	(0.013)	(0.014)	(0.016)
Parents' education level		0.013	0.003
		(0.019)	(0.019)
Income (std.)			0.119**
			(0.040)
Observations	431	375	347
Controls	Age	Age	Age
R^2	0.242	0.235	0.201
Adj. R^2	0.237	0.228	0.190

Note: OLS regressions with standard errors clustered at the household level (shown in parentheses). The dependent variable is digital skills, calculated as the unweighted average of all the survey items on digital skills, without standardization. *Male* is a binary indicator equal to 1 for male respondents. *Education Level* refers to years of formal education. *Parents' Education Level* refers to the years of education of the more highly educated parent. *Income (std. within ten-year age group)* denotes gross monthly income, standardized within ten-year age cohorts to account for income variation over the life cycle. The sample excludes individuals who are retired. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A9: Predictors of Digital Skills

	Dependent Variable: Digital Skills (std.)		
	(1)	(2)	(3)
Male	0.228***	0.246***	0.170**
	(0.052)	(0.053)	(0.057)
Abitur	0.320***	0.294***	0.180**
	(0.050)	(0.052)	(0.057)
Parents' Abitur		0.082	0.051
		(0.061)	(0.063)
Income (std.)			0.112***
			(0.028)
Observations	431	375	347
Controls	Age	Age	Age
R^2	0.195	0.194	0.191
Adj. R^2	0.190	0.186	0.180

Note: OLS regressions with standard errors clustered at the household level (shown in parentheses). The dependent variable is standardized digital skills, calculated as the unweighted average of all the survey items on digital skills, each standardized across participants. *Male* is a binary indicator equal to 1 for male respondents. *Abitur* is a binary dummy and indicates whether the person holds Abitur degree or not. *Parents' Abitur* is a binary dummy and indicates if at least one parent has an Abitur degree. *Income (std. within ten-year age group)* denotes gross monthly income standardized within ten-year age cohorts to account for income variation over the life cycle. The sample excludes individuals who are retired. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Predictors of Digital Skills Alternative Income Measures

	Dependent Variable: Digital Skills (std.)	
	(1)	(2)
Male	0.165**	0.213***
	(0.055)	(0.052)
Education	0.049***	0.066***
	(0.011)	(0.010)
Parents' education	0.003	0.008
	(0.013)	(0.012)
Income (std. full sample)	0.084**	
	(0.029)	
Household net income (std.)		0.032
		(0.027)
Observations	365	411
Controls	Age	Age
R^2	0.210	0.238
Adj.R^2	0.199	0.229

Note: OLS regressions with standard errors clustered at the household level (shown in parentheses). The dependent variable is standardized digital skills, calculated as the unweighted average of all survey items on digital skills, each standardized across participants. *Male* is a binary indicator equal to 1 for male respondents. *Education* refers to years of formal education. *Parents' education* refers to the years of education of the more highly educated parent. *Income (std. full sample)* refers to gross monthly income, standardized across all participants in the full sample. *Household net income (std.)* refers to net household income, standardized across all participants. The sample excludes individuals who are retired. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A11: Descriptive Statistics of Digital Skill Items (Full Sample)

Item	Mean	Median	SD
Aggregate	3.20	4	1.51
Technical_and_Operational_Skills_1	3.09	3	1.53
Technical_and_Operational_Skills_2	3.71	5	1.67
Technical_and_Operational_Skills_3	4.06	5	1.48
Technical_and_Operational_Skills_4	3.42	4	1.69
Technical_and_Operational_Skills_5	3.15	4	1.71
Technical_and_Operational_Skills_6	3.31	4	1.63
Programming_Skills_1	1.47	1	1.05
Information_Processing_and_Navigation_Skills_1	3.41	4	1.53
Information_Processing_and_Navigation_Skills_2	3.87	4	1.46
Information_Processing_and_Navigation_Skills_3	3.42	4	1.44
Information_Processing_and_Navigation_Skills_4	2.94	3	1.47
Information_Processing_and_Navigation_Skills_5	3.02	3	1.56
Information_Processing_and_Navigation_Skills_6	2.76	3	1.42

Notes: This table reports descriptive statistics for the full sample (N = 1,001) of thirteen digital skill survey items, which are measured on a scale from 1 to 5. Higher values indicate higher proficiency levels. The "Average" row shows the mean, median, and standard deviation computed from the unweighted average of all thirteen items.

Table A12: Descriptive Statistics of Digital Skill Items (Non-Retired)

Item	Mean	Median	SD
Aggregate	3.72	4	1.26
Technical_and_Operational_Skills_1	3.66	4	1.32
Technical_and_Operational_Skills_2	4.41	5	1.15
Technical_and_Operational_Skills_3	4.59	5	0.94
Technical_and_Operational_Skills_4	3.98	5	1.41
Technical_and_Operational_Skills_5	3.67	4	1.55
Technical_and_Operational_Skills_6	3.85	4	1.38
Programming_Skills_1	1.67	1	1.23
Information_Processing_and_Navigation_Skills_1	3.93	4	1.21
Information_Processing_and_Navigation_Skills_2	4.38	5	1.02
Information_Processing_and_Navigation_Skills_3	3.88	4	1.15
Information_Processing_and_Navigation_Skills_4	3.49	4	1.29
Information_Processing_and_Navigation_Skills_5	3.59	4	1.38
Information_Processing_and_Navigation_Skills_6	3.31	3	1.29

Notes: This table reports descriptive statistics for the sample excluding retired individuals (N = 539) of thirteen digital skill survey items, which are measured on a scale from 1 to 5. Higher values indicate higher proficiency levels. The "Average" row shows the mean, median, and standard deviation computed from the unweighted average of all thirteen items.

Table A13: Beliefs About Relative Skills Advantage Binary Education Measure

	DV: Belief About Relative Digital Skill Advantage (Std.)			
	(1)	(2)	(3)	(4)
	Population	Population	Age Group	Age Group
Male	0.325***	0.212*	0.453***	0.361***
	(0.095)	(0.090)	(0.094)	(0.089)
Abitur	0.353***	0.253***	0.324***	0.241**
	(0.103)	(0.094)	(0.112)	(0.109)
Parents' Abitur	0.143	0.120	0.104	0.090
	(0.103)	(0.093)	(0.114)	(0.113)
Income	0.053	-0.011	0.094	0.040
	(0.051)	(0.048)	(0.059)	(0.060)
Digital Skills (std.)		0.633***		0.501***
		(0.085)		(0.097)
Observations	338	338	340	340
Controls	Age	Age	Age	Age
R^2	0.239	0.358	0.134	0.207
Adj. R^2	0.228	0.346	0.121	0.193

Notes: OLS estimates with standard errors clustered at the household level (in parentheses). Dependent variables of Models (1) and (2) are standardized measures of subjective skill ranking relative to the general population. Dependent variables of Models (3) and (4) are standardized measures of subjective skill ranking relative to others in the same age group. *Male* is a binary indicator equal to 1 for male respondents. *Abitur* is a binary indicator equal to 1 if the respondent holds an Abitur (university-qualifying diploma). *Parents' Abitur* is a binary indicator equal to 1 if at least one partent holds an Abitur degree. *Income (Std.)* is gross monthly income standardized within ten-year age cohorts. *Digital Skills (std.)* is a standardized index of digital skill levels. The sample excludes retired individuals. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A14: Beliefs About Relative Skills Advantage Income (std. based on full sample)

	DV: Belief About Relative Digital Skill Advantage (Std.)			
	(1) Population	(2) Population	(3) Age Group	(4) Age Group
Male	0.345***	0.234***	0.452***	0.359***
	(0.090)	(0.088)	(0.093)	(0.088)
Education	0.083***	0.052***	0.077***	0.052***
	(0.019)	(0.018)	(0.020)	(0.019)
Parents' education level	0.029	0.030	0.007	0.008
	(0.022)	(0.021)	(0.024)	(0.024)
Income (std., full sample)	0.014	-0.025	0.050	0.009
	(0.055)	(0.051)	(0.063)	(0.063)
Digital skills (std.)		0.615***		0.527***
		(0.085)		(0.095)
Observations	344	344	348	348
Controls	Age	Age	Age	Age
R^2	0.258	0.366	0.140	0.215
Adj. R^2	0.247	0.354	0.129	0.202

Notes: OLS estimates with standard errors clustered at the household level (in parentheses). Dependent variables of Models (1) and (2) are standardized measures of subjective skill ranking relative to the general population. Dependent variables of Models (3) and (4) are standardized measures of subjective skill ranking relative to others in the same age group. Survey question: : What percentage of the German general population [those in your age group in Germany] do you believe has weaker digital skills than you? *Male* is a binary indicator equal to 1 for male respondents. *Education level* refers to years of formal education. *Parents' education level* refers to the years of education of the more highly educated parent. *Income (std. across the full sample)* denotes gross monthly income, standardized based on the full sample of the general population. All models control for age. The sample excludes retired individuals. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A15: Beliefs About Relative Skills Advantage Household Income

	DV: Belief About Relative Digital Skill Advantage (Std.)			
	(1) Population	(2) Population	(3) Age Group	(4) Age Group
Male	0.369***	0.227***	0.523***	0.400***
	(0.079)	(0.077)	(0.086)	(0.081)
Education	0.090***	0.048***	0.083***	0.048***
	(0.017)	(0.016)	(0.018)	(0.018)
Parents' education level	0.037	0.035	0.018	0.016
	(0.023)	(0.021)	(0.023)	(0.023)
Household net income (std.)	0.008	-0.004	0.050	0.039
	(0.041)	(0.038)	(0.044)	(0.042)
Digital skills (std.)		0.646***		0.557***
		(0.075)		(0.083)
Observations	385	385	390	390
Controls	Age	Age	Age	Age
R^2	0.300	0.419	0.159	0.248
Adj. R^2	0.291	0.410	0.148	0.236

Notes: OLS estimates with standard errors clustered at the household level (in parentheses). Dependent variables of Models (1) and (2) are standardized measures of subjective skill ranking relative to the general population. Dependent variables of Models (3) and (4) are standardized measures of subjective skill ranking relative to others in the same age group. Survey question: What percentage of the German general population [those in your age group in Germany] do you believe has weaker digital skills than you? *Male* is a binary indicator equal to 1 for male respondents. *Education level* refers to years of formal education. *Parents' education level* refers to the years of education of the more highly educated parent. *Household net income (std. across the full sample)* denotes monthly net household income, standardized based on the full sample. All models control for age. The sample excludes retired individuals. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A16: Miscalibration of Beliefs by Gender

	(1)	(2)	(3)
Intercept	-13.128^{***}	-13.657^{***}	-12.514^{***}
	(1.237)	(1.242)	(1.549)
Skills	-0.620^{***}	-0.752^{***}	-0.780^{***}
	(0.042)	(0.057)	(0.067)
Male	4.167^{*}	4.068^{*}	1.170
	(1.743)	(1.721)	(2.256)
Male × Skills		0.239^{***}	0.279^{***}
		(0.073)	(0.081)
Skills^2			-0.002
			(0.002)
Male × Skills^2			0.005^{*}
			(0.003)
Observations	481	481	481
Controls	Age (centered)	Age (centered)	Age (centered)
Adj. R^2	0.367	0.381	0.382

Note: OLS regressions with SE clustered at the household level (in parentheses). The dependent variable is the miscalibration of beliefs (estimated digital skill rank (percentile) minus actual rank based on the representative SOEP sample). Positive values indicate overestimation; negative underestimation. *Skills* refers to actual skill rank, centered by subtracting the sample mean to reduce multicollinearity with its square. *Skills*^{2} is the square of the centered skill rank. *Male* is a binary indicator equal to 1 for male respondents. All models control for centered age. Centering sets the intercept at the sample mean of age. The sample excludes retirees. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A17: Miscalibration of Beliefs by Education

	(1)	(2)	(3)
Intercept	-13.304^{***}	-13.468^{***}	-14.383^{***}
	(1.246)	(1.260)	(1.592)
Skills	-0.603^{***}	-0.646^{***}	-0.627^{***}
	(0.042)	(0.050)	(0.059)
Abitur	6.702^{***}	5.921^{**}	6.873^{**}
	(1.930)	(1.900)	(2.540)
Abitur × Skills		0.160^{*}	0.141
		(0.088)	(0.095)
Skills^2			0.002
			(0.002)
Abitur × Skills^2			−0.002
			(0.004)
Observations	408	408	408
Controls	Age (centered)	Age (centered)	Age (centered)
Adj. R^2	0.356	0.360	0.358

Note: OLS regressions with SE clustered at the household level (in parentheses). The dependent variable is the miscalibration of beliefs (estimated digital skill rank (percentile) minus actual rank based on the representative SOEP sample). Positive values indicate overestimation; negative underestimation. *Skills* refers to actual skill rank, centered by subtracting the sample mean to reduce multicollinearity with its square. *Skills*^{2} is the square of the centered skill rank. *Abitur* is a binary indicator equal to 1 for respondents who hold the *Abitur* degree, the highest secondary school qualification in Germany. All models control for centered age. Centering sets the intercept at the sample mean of age. The sample excludes retirees. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A18: Beliefs About Whether Digitalization is a Chance or Risk for Society

	DV: Digitalization is a Chance for Society (Belief)			
	(1)	(2)	(3)	(4)
Male	0.124	0.127	0.117	0.102
	(0.078)	(0.094)	(0.095)	(0.095)
Education	0.021	0.020	0.021	0.017
	(0.014)	(0.017)	(0.017)	(0.017)
Income		0.026	0.010	0.001
		(0.053)	(0.051)	(0.052)
Parent education			-0.028	-0.028
			(0.020)	(0.021)
Digital skills (std.)				0.115
				(0.096)
Observations	430	375	365	365
Adj. R^2	0.018	0.015	0.009	0.010
Controls	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control

Note: OLS regressions with standard errors clustered at the household level (in parentheses). The dependent variable is standardized belief that digitalization is a chance (rather than a risk) for society. *Male* is a binary indicator equal to 1 for male respondents. *Education* is measured in years of completed formal education. *Income* is standardized within ten-year age cohorts. *Parent education* is measured in years of completed formal education of the more educated parent. *Digital skills (std.)* is a standardized measure of actual digital skills. All models control for age and locus of control, measured by agreement with: The outcome of my life is within my control. The sample excludes individuals in retirement. Standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A19: Beliefs About Whether Digitalization is a Chance or Risk for Leisure Time

	DV: Digitalization is a Chance for Leisure Time (Belief)			
	(1)	(2)	(3)	(4)
Male	0.068	0.070	0.058	0.038
	(0.067)	(0.077)	(0.076)	(0.077)
Education	0.010	0.004	0.001	-0.005
	(0.014)	(0.016)	(0.015)	(0.015)
Income		-0.005	-0.022	-0.034
		(0.045)	(0.045)	(0.047)
Parent education			-0.010	-0.010
			(0.020)	(0.020)
Digital skills (std.)				0.144*
				(0.084)
Observations	430	375	365	365
Adj. R^2	0.002	-0.006	-0.010	-0.002
Controls	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control

Note: OLS regressions with standard errors clustered at the household level (in parentheses). The dependent variable is standardized belief that digitalization is a chance (rather than a risk) for leisure time. *Male* is a binary indicator equal to 1 for male respondents. *Education* is measured in years of completed formal education. *Income* is standardized within ten-year age cohorts. *Parent education* is measured in years of completed formal education of the more educated parent. *Digital skills (std.)* is a standardized measure of actual digital skills. All models control for age and locus of control, measured by agreement with: The outcome of my life is within my control. The sample excludes individuals in retirement. Standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A20: Beliefs About Whether Digitalization is a Chance or Risk for Relationships with Friends and Family

	DV: Digitalization is a Chance for Relationships (Belief)			
	(1)	(2)	(3)	(4)
Male	0.008	0.051	0.026	0.011
	(0.070)	(0.081)	(0.081)	(0.082)
Education	-0.002	-0.000	0.002	-0.002
	(0.015)	(0.018)	(0.017)	(0.018)
Income		-0.022	-0.051	-0.060
		(0.047)	(0.044)	(0.046)
Parent education			-0.033*	-0.033*
			(0.019)	(0.019)
Digital skills (std.)				0.113
				(0.084)
Observations	430	375	365	365
Adj. R^2	-0.002	-0.004	0.002	0.005
Controls	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control	Age, Locus of Control

Note: OLS regressions with standard errors clustered at the household level (in parentheses). The dependent variable is standardized belief that digitalization is a chance (rather than a risk) for relationships with friends and family. *Male* is a binary indicator equal to 1 for male respondents. *Education* is measured in years of completed formal education. *Income* is standardized within ten-year age cohorts. *Parent education* is measured in years of completed formal education of the more educated parent. *Digital skills (std.)* is a standardized measure of actual digital skills. All models control for age and locus of control, measured by agreement with: The outcome of my life is within my control. The sample excludes individuals in retirement. Standard errors in parentheses. ${}^{*}p < 0.1$, ${}^{**}p < 0.05$, ${}^{***}p < 0.01$.

Table A21: Digital Skills - Item Wording (German)

Questionnaire Items (examples) (13 items in total)
Anwendungs- und technologische Fähigkeiten:
Ich weiSS, wie man Datenschutzeinstellungen anpasst.
Ich weiSS, wie man den Standort auf mobilen Geräten ausschaltet.
Ich weiSS, wie man ein Gerät schützt (z.B. mit PIN, Bildschirmmuster, Fingerabdruck oder Gesichtserkennung).
Ich weiSS, wie man Fotos, Dokumente oder andere Dateien in einer Cloud speichert (z.B. Google Drive, iCloud).
Ich weiSS, wie man privates Surfen einstellt.
Ich weiSS, wie man unerwünschte Pop-Up-Nachrichten oder Anzeigen blockiert.
Programmieren:
Ich kann eine Programmiersprache (z.B. XML, Python, Java, C++) anwenden.
Information und Navigation:
Ich weiSS, welche Stichwörter man am besten bei einer Internet-Suche wählt/eingibt.
Ich weiSS, wie ich eine Webseite wiederfinde, die ich bereits besucht habe.
Ich weiSS, wie ich Informationen auf einer Webseite finde, egal wie sie aufgebaut ist.
Ich weiSS, wie ich herausfinde, ob eine Webseite vertrauenswürdig ist.
Ich weiSS, wie man erweiterte Suchfunktionen in Suchmaschinen verwendet.
Ich weiSS, wie ich überprüfen kann, ob die im Internet gefundenen Informationen wahr sind.

Skala: Trifft überhaupt nicht zu, Trifft eher nicht zu, Trifft teils zu und teils nicht zu, Trifft eher zu, Trifft voll und ganz zu, Ich weiSS nicht, was damit gemeint ist, Das möchte ich nicht sagen

Table A22: Digital Skills - Item Wording (English)

Questionnaire Items (examples) (13 items in total)
Technical and operational skills:
I know how to adjust privacy settings.
I know how to turn off the location settings on mobile devices.
I know how to protect a device (e.g., with a PIN).
I know how to store photos, documents, or other files in the cloud (e.g., Google Drive, iCloud).
I know how to use private browsing (e.g., incognito mode).
I know how to block unwanted pop-up messages or ads.
Programming:
I know how to use programming languages (e.g., XML, Python, Java, C++).
Information navigation and processing:
I know how to choose the best keywords for online searches.
I know how to find a website I have visited before.
I know how to find information on a website, no matter how it is designed.
I know how to figure out if a website can be trusted.
I know how to use advanced search functions in search engines.
I know how to check if the information I find online is true.

Scale: strongly disagree, disagree, neutral, agree, strongly agree, I don’t know, Prefer not to say

List of Figures

List of Tables

Bibliography

Acemoglu, D. (2002): “Technical Change, Inequality, and the Labor Market,” *Journal of Economic Literature*, 40, 7–72.

——— (2021a): “Harms of AI. Working Paper.” .

——— (2021b): *Redesigning AI*, MIT Press.

Acemoglu, D. and D. Autor (2011): “Skills, tasks and technologies: Implications for employment and earnings,” in *Handbook of labor economics*, Elsevier, vol. 4, 1043–1171.

Acemoglu, D., D. H. Autor, J. Hazell, and P. Restrepo (2022): “Artificial intelligence and jobs: Evidence from online vacancies,” *Journal of Labor Economics*, 40, 293–340.

Acemoglu, D. and P. Restrepo (2018): “Low-skill and High-skill Automation,” *Journal of Human Capital*, 12, 204–232.

——— (2020): “Robots and Jobs: Evidence from US Labor Markets,” *Journal of Political Economy*, 128, 2188–2244.

Adda, J., C. Dustmann, and J.-S. Görlach (2022): “The dynamics of return migration, human capital accumulation, and wage assimilation,” *The Review of Economic Studies*, 89, 2841–2871.

Adoma, A. F., N.-M. Henry, and W. Chen (2020): “Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition,” in *2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, IEEE, 117–121.

Agan, A. Y., D. Davenport, J. Ludwig, and S. Mullainathan (2023): “Automating automaticity: How the context of human choice affects the extent of algorithmic bias,” Tech. rep., National Bureau of Economic Research.

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023): “Combining human expertise with artificial intelligence: Experimental evidence from radiology,” Tech. rep., National Bureau of Economic Research.

Aksoy, C. G., B. Özcan, and J. Philipp (2021): “Robots and the gender pay gap in Europe,” *European Economic Review*, 134, 103693.

Albright, A. (2019): “If you give a judge a risk score: evidence from Kentucky bail decisions,” .

Alekseeva, L., J. Azar, M. Gine, S. Samila, and B. Taska (2021): “The demand for AI skills in the labor market,” *Labour Economics*, 71, 293 340.

Allen, R. and P. Choudhury (2022): “Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion,” *Organization Science*, 33, 149–169.

Alon-Barkat, S. and M. Busuioc (2023): “Human–AI interactions in public sector decision making:automation bias and selective adherence to algorithmic advice,” *Journal of Public Administration Research and Theory*, 33, 153–169.

Ambuehl, S. and S. Li (2018): “Belief updating and the demand for information,” *Games and Economic Behavior*, 109, 21–39.

Anderson, L. R., R. G. Fryer, and C. A. Holt (2006): “Discrimination: experimental evidence from psychology and economics,” *Handbook on the Economics of Discrimination.*

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022): “Subjective models of the macroeconomy: Evidence from experts and representative samples,” *The Review of Economic Studies*, 89, 2958–2991.

Andres, M., L. Bruttel, and J. Friedrichsen (2023): "How communication makes the difference between a cartel and tacit collusion: A machine learning approach," *European Economic Review*, 152, 104331.

Angelova, V., W. S. Dobbie, and C. Yang (2023): "Algorithmic recommendations and human discretion," Tech. rep., National Bureau of Economic Research.

Arnesen, S., T. S. Broderstad, J. S. Fishkin, M. P. Johannesson, and A. Siu (2024): "Knowledge and support for AI in the public sector: a deliberative poll experiment," *AI & SOCIETY*, 1–17.

Arnold, D., W. Dobbie, and P. Hull (2021): "Measuring racial discrimination in algorithms," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 111, 49–54.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton, NJ: Princeton University Press, 3–33.

Ash, E. and S. Hansen (2023): "Text algorithms in economics," *Annual Review of Economics*, 15, 659–688.

Ash, E., S. Hansen, C. Marangon, and Y. Muvdi (2025): "Large Language Models in Economics," Technical report, New Palgrave Dictionary of Economics.

Athey, S. C., K. A. Bryan, and J. S. Gans (2020): "The allocation of decision authority to human and artificial intelligence," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 110, 80–84.

Autor, D., C. Chin, A. Salomons, and B. Seegmiller (2024): "New frontiers: The origins and content of new work, 1940–2018," *The Quarterly Journal of Economics*, 139, 1399–1465.

Autor, D., F. Levy, and R. Murnane (2003): "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Avery, C. and P. Zemsky (1998): "Multidimensional uncertainty and herd behavior in financial markets," *American economic review*, 724–748.

Avery, M., A. Leibbrandt, and J. Vecci (2024): "Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech," .

Banerjee, A. V. (1992): "A simple model of herd behavior," *The quarterly journal of economics*, 107, 797–817.

Bansak, K. and E. Paulson (2024): "Public attitudes on performance for algorithmic and human decision-makers," *PNAS nexus*, 3, pgae520.

Barabas, J. (2004): "How deliberation affects policy opinions," *American political science review*, 98, 687–701.

Barber, B. M. and T. Odean (2001): "Boys will be boys: Gender, overconfidence, and common stock investment," *The quarterly journal of economics*, 116, 261–292.

Barocas, S., M. Hardt, and A. Narayanan (2023): *Fairness and machine learning: Limitations and opportunities*, MIT press.

Barocas, S. and A. D. Selbst (2016): "Big Data's Disparate Impact," *California Law Review*, 104.

Barron, K., R. Ditlmann, S. Gehrig, and S. Schweighofer-Kodritsch (2024): "Explicit and implicit belief-based gender discrimination: A hiring experiment," *Management Science*.

Becker, G. (1957): "S.(1971),The Economics of Discrimination," .

Becker, G. S. and N. Tomes (1979): "An equilibrium theory of the distribution of income and intergenerational mobility," *Journal of political Economy*, 87, 1153–1189.

Bénabou, R. and J. Tirole (2002): "Self-confidence and personal motivation," *The quarterly journal of economics*, 117, 871–915.

Benjamin, D. J. (2019): "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2, 69–186.

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018): "quanteda: An R package for the quantitative analysis of textual data," *Journal of Open Source Software*, 3, 774–774.

Benyishay, A., M. Jones, F. Kondylis, and A. M. Mobarak (2020): "Gender gaps in technology diffusion," *Journal of development economics*, 143, 102380.

Bertrand, M. (2020): "Gender in the twenty-first century," in *AEA Papers and proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 110, 1–24.

Bertrand, M. and E. Duflo (2017): "Field experiments on discrimination," *Handbook of economic field experiments*, 1, 309–393.

Bertrand, M. and S. Mullainathan (2001): "Do people mean what they say? Implications for subjective survey data," *American Economic Review*, 91, 67–72.

Bick, A., A. Blandin, and D. J. Deming (2024): "The Rapid Adoption of Generative AI," NBER Working Paper No. 32966.

Bikhchandani, S., D. Hirshleifer, O. Tamuz, and I. Welch (2024): "Information cascades and social learning," *Journal of Economic Literature*, 62, 1040–1093.

Blau, F. D. and L. M. Kahn (2017): "The gender wage gap: Extent, trends, and explanations," *Journal of economic literature*, 55, 789–865.

Bogert, E., A. Schecter, and R. T. Watson (2021): "Humans rely more on algorithms than social influence as a task becomes more difficult," *Scientific reports*, 11, 8028.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2025): "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 1–16.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The dynamics of discrimination: Theory and evidence," *American economic review*, 109, 3395–3436.

Bojić, L., O. Zagovora, A. Zelenkauskaite, V. Vuković, M. Čabarkapa, S. Veseljević Jerković, and A. Jovančević (2025): "Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm," *Scientific reports*, 15, 11477.

Bonnefon, J.-F., A. Shariff, and I. Rahwan (2016): "The social dilemma of autonomous vehicles," *Science*, 352, 1573–1576.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

——— (2019): "Beliefs about gender," *American Economic Review*, 109, 739–773.

Brandts, J., D. J. Cooper, and C. Rott (2019): "21. Communication in laboratory experiments," *Handbook of research methods and applications in experimental economics*, 401.

Breda, T., J. Grenet, M. Monnet, and C. Van Effenterre (2023): "How effective are female role models in steering girls towards STEM? Evidence from French high schools," *The Economic Journal*, 133, 1773–1809.

Brunnermeier, M. K. and J. A. Parker (2005): "Optimal expectations," *American Economic Review*, 95, 1092–1118.

Brynjolfsson, E., D. Li, and L. Raymond (2025): "Generative AI at work," *The Quarterly Journal of Economics*, qjae044.

Bursztyn, L., A. Rao, C. Roth, and D. Yanagizawa-Drott (2023): "Opinions as facts," *The Review of Economic Studies*, 90, 1832–1864.

Burton, J. W., M.-K. Stein, and T. B. Jensen (2020): "A systematic review of algorithm aversion in augmented decision making," *Journal of behavioral decision making*, 33, 220–239.

Buser, T., M. Niederle, and H. Oosterbeek (2014): "Gender, competitiveness, and career choices," *The quarterly journal of economics*, 129, 1409–1447.

Campos-Mercade, P. and F. Mengel (2024): "Non-Bayesian statistical discrimination," *Management Science*, 70, 2549–2567.

Capraro, V., A. Lentsch, D. Acemoglu, S. Akgun, A. Akhmedova, E. Bilancini, J.-F. Bonnefon, P. Brañas-Garza, L. Butera, K. M. Douglas, et al. (2024): "The impact of generative artificial intelligence on socioeconomic inequalities and policy making," *PNAS nexus*, 3, pgae191.

Carlana, M. (2019): "Implicit stereotypes: Evidence from teachers gender bias," *The Quarterly Journal of Economics*, 134, 1163–1224.

Carvajal, D., C. Franco, and S. Isaksson (2024): "Will Artificial Intelligence Get in the Way of Achieving Gender Equality?" NHH Dept. of Economics Discussion Paper No. 03.

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

Cazzaniga, M., A. Panton, L. Li, C. Pizzinelli, and M. M. Tavares (2025): "A Gender Lens on Labor Market Exposure to AI," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 115, 56–61.

Celiktutan, B., R. Cadario, and C. K. Morewedge (2024): "People see more of their biases in algorithms," *Proceedings of the National Academy of Sciences*, 121, e2317602121.

Chen, B., Y. Hermstrüwer, P. Langenbach, A. Stremitzer, and K. Tobia (2025): "Mitigating the Judicial Human-AI Fairness Gap," *Center for Law & Economics Working Paper Series*, 7.

Chen, B. M., A. Stremitzer, and K. Tobia (2022): "Having your day in robot court," *Harv. JL & Tech.*, 36, 127.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree - An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014): "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The quarterly journal of economics*, 129, 1553–1623.

Chopra, F., I. Haaland, and C. Roth (2024): "The demand for news: Accuracy concerns versus belief confirmation motives," *The Economic Journal*, 134, 1806–1834.

Coffman, K., M. R. Collis, and L. Kulkarni (2024a): "Stereotypes and belief updating," *Journal of the European Economic Association*, 22, 1011–1054.

Coffman, K., C. B. Flikkema, and O. Shurchkov (2021a): "Gender stereotypes in deliberation and team decisions," *Games and Economic Behavior*, 129, 329–349.

Coffman, K. B. (2014): "Evidence on self-stereotyping and the contribution of ideas," *The Quarterly Journal of Economics*, 129, 1625–1660.

Coffman, K. B., M. R. Collis, and L. Kulkarni (2024b): "Whether to Apply," *Management Science*, 70, 4649–4669.

Coffman, K. B., C. L. Exley, and M. Niederle (2021b): "The role of beliefs in driving gender discrimination," *Management Science*, 67, 3551–3569.

Cohen, J., K. M. Ericson, D. Laibson, and J. M. White (2020): "Measuring time preferences," *Journal of Economic Literature*, 58, 299–347.

Conlon, J. J., M. Mani, G. Rao, M. W. Ridley, and F. Schilbach (2022): "Not learning from others," Tech. rep., National Bureau of Economic Research.

Corak, M. (2013): "Income inequality, equality of opportunity, and intergenerational mobility," *Journal of Economic Perspectives*, 27, 79–102.

Corbett-Davies, S., J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel (2023): "The measure and mismeasure of fairness," *The Journal of Machine Learning Research*, 24, 14730–14846.

Corgnet, B., M. DeSantis, and D. Porter (2024): "Lets chat when communication promotes efficiency in experimental asset markets," *Management Science*, 70, 6550–6568.

Coutts, A. (2019): "Good news and bad news are still news: Experimental evidence on belief updating," *Experimental Economics*, 22, 369–395.

Cowgill, B. (2020): "Bias and productivity in humans and algorithms: Theory and evidence from resume screening," *Columbia Business School, Columbia University*, 29.

Cowgill, B., F. DellAcqua, and S. Matz (2020): "The managerial effects of algorithmic fairness activism," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 110, 85–90.

Cowgill, B. and C. E. Tucker (2020): "Economics, fairness and algorithmic bias," *in preparation for: Journal of Economic Perspectives.*

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 112, 2851–2883.

Dargnies, M.-P., R. Hakimov, and D. Kübler (2024): "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence," *Management Science.*

Das, S., C. M. Kuhnen, and S. Nagel (2020): "Socioeconomic status and macroeconomic expectations," *The Review of Financial Studies*, 33, 395–432.

Davenport, D. (2023): "Discriminatory Discretion: Theory and evidence from use of pretrial algorithms," *Available at SSRN.*

De-Arteaga, M., V. Jeanselme, A. Dubrawski, and A. Chouldechova (2025): "Leveraging expert consistency to improve algorithmic decision support," *Management Science.*

De Freitas, J., S. Agarwal, B. Schmitt, and N. Haslam (2023): "Psychological factors underlying attitudes toward AI tools," *Nature Human Behaviour*, 7, 1845–1854.

Delavande, A. and B. Zafar (2018): "Information and anti-American attitudes," *Journal of Economic Behavior & Organization*, 149, 1–31.

DellaVigna, S. and M. Gentzkow (2010): "Persuasion: empirical evidence," *Annu. Rev. Econ.*, 2, 643–669.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dietvorst, B. J., J. P. Simmons, and C. Massey (2015): "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, 144, 114.

——— (2018): "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management science*, 64, 1155–1170.

DAcunto, F., D. Hoang, M. Paloviita, and M. Weber (2019): "Cognitive abilities and inflation expectations," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 109, 562–566.

Dacunto, F., D. Hoang, M. Paloviita, and M. Weber (2023): "IQ, expectations, and choice," *The Review of Economic Studies*, 90, 2292–2325.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024): "GPTs are GPTs: Labor market impact potential of LLMs," *Science*, 384, 1306–1308.

ENKE, B. AND T. GRAEBER (2023): "Cognitive uncertainty," *The Quarterly Journal of Economics*, 138, 2021–2067.

ENKE, B. AND F. ZIMMERMANN (2019): "Correlation neglect in belief formation," *The review of economic studies*, 86, 313–332.

ERIKSON, R. AND J. H. GOLDTHORPE (2002): "Intergenerational inequality: A sociological perspective," *Journal of Economic Perspectives*, 16, 31–44.

ESPONDA, I., R. OPREA, AND S. YUKSEL (2023): "Seeing what is representative," *The Quarterly Journal of Economics*, 138, 2607–2657.

EXLEY, C. L. AND J. B. KESSLER (2022): "The gender gap in self-promotion," *The Quarterly Journal of Economics*, 137, 1345–1381.

EXLEY, C. L. AND K. NIELSEN (2024): "The gender gap in confidence: Expected but not accounted for," *American Economic Review*, 114, 851–885.

FAFCHAMPS, M., A. ISLAM, D. PAKRASHI, AND D. TOMMASI (2024): "Diffusion in social networks: Experimental evidence on information sharing vs persuasion," Tech. rep., National Bureau of Economic Research.

FERRARIO, B. AND S. STANTCHEVA (2022): "Eliciting peoples first-order concerns: Text analysis of open-ended survey questions," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 112, 163–169.

FISCHBACHER, U., L. NEYSE, D. RICHTER, AND C. SCHRÖDER (2024): "Adding household surveys to the behavioral economics toolbox: insights from the SOEP innovation sample," *Journal of the Economic Science Association*, 10, 136–151.

FISHKIN, J. S. (2018): *Democracy when the people are thinking: Revitalizing our politics through public deliberation*, Oxford University Press.

FREY, C. B. AND M. A. OSBORNE (2017): "The future of employment: How susceptible are jobs to computerisation?" *Technological forecasting and social change*, 114, 254–280.

FRYER, R. G., S. D. LEVITT, AND J. A. LIST (2008): "Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study," *American Economic Review*, 98, 370–375.

FRYER JR, R. G., P. HARMS, AND M. O. JACKSON (2019): "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization," *Journal of the European Economic Association*, 17, 1470–1501.

GARCIA, D., J. TOLVANEN, AND A. K. WAGNER (2024): "Strategic Responses to Algorithmic Recommendations: Evidence from Hotel Pricing," *Management Science*.

GATHMANN, C., C. KAGERL, L. POHLAN, AND D. ROTH (2024): "The pandemic push: Digital technologies and workforce adjustments," *Labour Economics*, 89, 102541.

GENICOT, G. AND D. RAY (2017): "Aspirations and inequality," *Econometrica*, 85, 489–519.

GENNAIOLI, N. AND A. SHLEIFER (2010): "What comes to mind," *The Quarterly journal of economics*, 125, 1399–1433.

GENTZKOW, M., B. KELLY, AND M. TADDY (2019): "Text as data," *Journal of Economic Literature*, 57, 535–574.

GENTZKOW, M. AND J. M. SHAPIRO (2006): "Media bias and reputation," *Journal of political Economy*, 114, 280–316.

GESK, T. S. AND M. LEYER (2022): "Artificial intelligence in public services: When and why citizens accept its usage," *Government Information Quarterly*, 39, 101704.

GILLIS, T., B. MCLAUGHLIN, AND J. SPIESS (2021): "On the fairness of machine-assisted human decisions," *arXiv preprint arXiv:2110.15310*.

Gillis, T. B. and J. L. Spiess (2019): “Big data and discrimination,” *The University of Chicago Law Review*, 86, 459–488.

Glickman, M. and T. Sharot (2025): “How human–AI feedback loops alter human perceptual, emotional and social judgements,” *Nature Human Behaviour*, 9, 345–359.

Goeree, J. K. and L. Yariv (2011a): “An experimental study of collective deliberation,” *Econometrica*, 79, 893–921.

——— (2011b): “An Experimental Study of Collective Deliberation,” *Econometrica*, 79, 893–921.

Goldin, C. (1994): “Understanding the gender gap: An economic history of American women,” *Equal employment opportunity: labor market discrimination and public policy*, 17–26.

Griffin, D. and A. Tversky (1992): “The weighing of evidence and the determinants of confidence,” *Cognitive psychology*, 24, 411–435.

Grunewald, A., V. Klockmann, A. von Schenk, and F. von Siemens (2024): “Are biases contagious? The influence of communication on motivated beliefs,” Tech. rep., Würzburg Economic Papers.

Haaland, I., C. Roth, S. Stantcheva, and J. Wohlfart (2025): “Understanding economic behavior using open-ended survey data,” Tech. rep., ECONtribute Discussion Paper.

Haaland, I., C. Roth, and J. Wohlfart (2023): “Designing information provision experiments,” *Journal of economic literature*, 61, 3–40.

Hausladen, C. I., M. Fochmann, and P. Mohr (2024): “Predicting compliance: Leveraging chat data for supervised classification in experimental research,” *Journal of Behavioral and Experimental Economics*, 109, 102164.

Hausman, J. (2012): “Contingent valuation: from dubious to hopeless,” *Journal of economic perspectives*, 26, 43–56.

Heckman, J. J. (1976): “A life-cycle model of earnings, learning, and consumption,” *Journal of political economy*, 84, S9–S44.

——— (2011): “The economics of inequality: The value of early childhood education.” *American Educator*, 35, 31.

Hellman, D. (2020): “Measuring algorithmic fairness,” *Virginia Law Review*, 106, 811–866.

Helsper, E. J., L. S. Schneider, A. J. van Deursen, and E. van Laar (2020): “The youth Digital Skills Indicator: Report on the conceptualisation and development of the ySKILLS digital skills measure,” *KU Leuven, Leuven: ySKILLS.*

Holzer, H. and D. Neumark (2000): “Assessing affirmative action,” *Journal of Economic literature*, 38, 483–568.

Hughey, M. W. (2022): “Superposition strategies: How and why White people say contradictory things about race,” *Proceedings of the National Academy of Sciences*, 119, e2116306119.

Humlum, A. and E. Vestergaard (2025): “The unequal adoption of ChatGPT exacerbates existing inequalities among workers,” *Proceedings of the National Academy of Sciences*, 122, e2414972121.

Hüning, H., L. Mechtenberg, and S. Wang (2022a): “Detecting arguments and their positions in experimental communication data,” *Available at SSRN 4052402.*

——— (2022b): “Using Arguments to Persuade: Experimental Evidence,” *Available at SSRN 4244989.*

Iaryczower, M., X. Shi, and M. Shum (2018): “Can words get in the way? The effect of deliberation in collective decision making,” *Journal of Political Economy*, 126, 688–734.

Jackman, J. A., D. A. Gentile, N.-J. Cho, and Y. Park (2021): “Addressing the digital skills gap for future education,” *Nature Human Behaviour*, 5, 542–545.

Jäger, S., C. Roth, N. Roussille, and B. Schoefer (2024): “Worker Beliefs about Outside Options,” *The Quarterly Journal of Economics*, 139, 15051556.

Jussupow, E., I. Benbasat, and A. Heinzl (2020): "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion," .

Kahneman, D. and A. Tversky (1973): "On the psychology of prediction." *Psychological review*, 80, 237.

Kallus, N., X. Mao, and A. Zhou (2022): "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*, 68, 1959–1981.

Kalyani, A., N. Bloom, M. Carvalho, T. Hassan, J. Lerner, and A. Tahoun (2025): "The diffusion of new technologies," *The Quarterly Journal of Economics*, 140, 1299–1365.

Kamenica, E. (2019): "Bayesian persuasion and information design," *Annual Review of Economics*, 11, 249–272.

Kamenica, E. and M. Gentzkow (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

Kaniel, R., C. Massey, and D. T. Robinson (2010): "The Importance of Being an Optimist: Evidence from Labor Markets," NBER Working Paper No. 16328.

Karni, E. (2009): "A mechanism for eliciting probabilities," *Econometrica*, 77, 603–606.

Kawaguchi, K. (2021): "When will workers follow an algorithm? A field experiment with a retail business," *Management Science*, 67, 1670–1695.

Kearney, M. S. and P. B. Levine (2014): "Income inequality, social mobility, and the decision to drop out of high school," Tech. rep., National Bureau of Economic Research.

Keller, W. (2004): "International technology diffusion," *Journal of economic literature*, 42, 752–782.

Kelly, S., S.-A. Kaye, and O. Oviedo-Trespalacios (2023): "What factors contribute to the acceptance of artificial intelligence? A systematic review," *Telematics and Informatics*, 77, 101925.

Kennedy, R. P., P. D. Waggoner, and M. M. Ward (2022): "Trust in public policy algorithms," *The Journal of Politics*, 84, 1132–1148.

Kim, H., E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca (2024): "Decision authority and the returns to algorithms," *Strategic Management Journal*, 45, 619–648.

Kim, T. and W. Peng (2024): "Do we want AI judges? The acceptance of AI judges judicial decision-making on moral foundations," *AI & SOCIETY*, 1–14.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018a): "Human decisions and machine predictions," *The quarterly journal of economics*, 133, 237–293.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018b): "Algorithmic fairness," in *Aea papers and proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 108, 22–27.

Korinek, A. (2023): "Generative AI for economic research: Use cases and implications for economists," *Journal of Economic Literature*, 61, 1281–1317.

Lafont, C. (2023): "Democracy Without Shortcuts: An Institutional Approach to Democratic Legitimacy," in *Another Universalism: Seyla Benhabib and the Future of Critical Theory*, Columbia University Press, 151–165.

Lambrecht, A. and C. Tucker (2019): "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management science*, 65, 2966–2981.

——— (2024): "Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising," *Quantitative Marketing and Economics*, 1–31.

Lange, K.-R., M. Reccius, T. Schmidt, H. Müller, M. W. Roos, and C. Jentsch (2022): *Towards extracting collective economic narratives from texts*, 963, Ruhr Economic Papers.

Liu, M., X. Tang, S. Xia, S. Zhang, Y. Zhu, and Q. Meng (2023): "Algorithm aversion: Evidence from ridesharing drivers," *Management Science*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019): "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*.

Logg, J. M., J. A. Minson, and D. A. Moore (2019): "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, 151, 90–103.

LSE Media and Communications (2024): "ySKILLS: Youth Digital Skills, Digital Literacies, and Resilience," `https://www.lse.ac.uk/media-and-communications/research/research-projects/ySKILLS`.

Ludwig, J. and S. Mullainathan (2021): "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System," *Journal of Economic Perspectives*, 35, 71–96.

Malmendier, U. and G. Tate (2005): "CEO Overconfidence and Corporate Investment," *The Journal of Finance*, 60, 2661–2700.

Mann, R. P. (2020): "Collective decision-making by rational agents with differing preferences," *Proceedings of the National Academy of Sciences*, 117, 10388–10396.

Matthew, A., A. R. Miller, and C. Tucker (2024): "Algorithmic Bias and Historical Injustice: Race and Digital Profiling," Tech. rep., National Bureau of Economic Research.

McLaughlin, B. and J. Spiess (2022): "Algorithmic assistance with recommendation-dependent preferences," *arXiv preprint arXiv:2208.07626*.

Miller, A. R. and C. Segal (2012): "Does temporary affirmative action produce persistent effects? A study of black and female employment in law enforcement," *Review of Economics and Statistics*, 94, 1107–1125.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2021): "Algorithmic fairness: Choices, assumptions, and definitions," *Annual review of statistics and its application*, 8, 141–163.

Möbius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2022): "Managing self-confidence: Theory and experimental evidence," *Management Science*, 68, 7793–7817.

Mobius, M. M. and T. S. Rosenblat (2006): "Why beauty matters," *American Economic Review*, 96, 222–235.

Moreno, M. J., P. Ouss, and B. A. Ba (2025): "Officer-involved: The media language of police killings," *The Quarterly Journal of Economics*, 140, 1525–1580.

Morewedge, C. K., S. Mullainathan, H. F. Naushan, C. R. Sunstein, J. Kleinberg, M. Raghavan, and J. O. Ludwig (2023): "Human bias in algorithm design," *Nature Human Behaviour*, 7, 1822–1824.

Muehlheusser, G., T. Promann, A. Roider, and N. Wallmeier (2024): "Honesty of groups: Effects of size and gender composition," Tech. rep., IZA Discussion Papers.

Mueller, A. I., J. Spinnewijn, and G. Topa (2021): "Job seekers perceptions and employment prospects: Heterogeneity, duration dependence, and bias," *American Economic Review*, 111, 324–363.

Mui, P. and B. Schoefer (2025): "Reservation Raises: The Aggregate Labour Supply Curve at the Extensive Margin," *Review of Economic Studies*, 92, 442–475.

Narayanan, A. and S. Kapoor (2024): "AI snake oil: What artificial intelligence can do, what it cant, and how to tell the difference," in *AI Snake Oil*, Princeton University Press.

Neumark, D., I. Burn, and P. Button (2016): "Experimental age discrimination evidence and the Heckman critique," *American Economic Review*, 106, 303–308.

Ng, A. Y. (2004): "Feature selection, L 1 vs. L 2 regularization, and rotational invariance," in *Proceedings of the twenty-first international conference on Machine learning*, 78.

Niederle, M., C. Segal, and L. Vesterlund (2013): “How costly is diversity? Affirmative action in light of gender differences in competitiveness,” *Management Science*, 59, 1–16.

Nielsen, A. and A. Woemmel (2024): “Invisible Inequities: Confronting Age-Based Discrimination in Machine Learning Research and Applications,” in *Proceedings of the 2nd Workshop on Generative AI and Law (GenLaw 24), International Conference on Machine Learning (ICML)*, Vienna, Austria.

Noy, S. and W. Zhang (2023): “Experimental evidence on the productivity effects of generative artificial intelligence,” *Science*, 381, 187–192.

Nussberger, A.-M., L. Luo, L. E. Celis, and M. J. Crockett (2022): “Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence,” *Nature Communications*, 13, 5821.

Oreopoulos, P. (2011): “Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes,” *American Economic Journal: Economic Policy*, 3, 148–171.

Owen, S. (2023): “College major choice and beliefs about relative performance: An experimental intervention to understand gender gaps in STEM,” *Economics of Education Review*, 97, 102479.

Pethig, F. and J. Kroenung (2023): “Biased humans,(un) biased algorithms?” *Journal of Business Ethics*, 183, 637–652.

Phelps, E. S. (1972): “The Statistical Theory of Racism and Sexism,” *The American Economic Review*, 62, 659–661.

Promann, T., J.-P. Mayer, G. Muehlheusser, A. Roider, E. Tereschenko, and N. Wallmeier (2025): “chaTree: An oTree addon allowing face-to-face communication in online group experiments,” *Economics Letters*, 112349.

Puri, M. and D. T. Robinson (2007): “Optimism and economic choice,” *Journal of Financial Economics*, 86, 71–99.

Rabin, M. and J. L. Schrag (1999): “First impressions matter: A model of confirmatory bias,” *The quarterly journal of economics*, 114, 37–82.

Rabinovitch, H., D. V. Budescu, and Y. B. Meyer (2024): “Algorithms in selection decisions: Effective, but unappreciated,” *Journal of Behavioral Decision Making*, 37, e2368.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020): “An Economic Perspective on Algorithmic Fairness,” *AEA Papers and Proceedings*, 110, 91–95.

Reich, T., A. Kaju, and S. J. Maglio (2023): “How to overcome algorithm aversion: Learning from mistakes,” *Journal of Consumer Psychology*, 33, 285–302.

Reuben, E., P. Sapienza, and L. Zingales (2014): “How stereotypes impair womens careers in science,” *Proceedings of the National Academy of Sciences*, 111, 4403–4408.

Richter, D. and J. Schupp (2012): “SOEP Innovation Sample (SOEP-IS) Description, Structure and Documentation,” SOEPpapers on Multidisciplinary Panel Data Research No. 463.

Roussille, N. (2024): “The role of the ask gap in gender pay inequality,” *The Quarterly Journal of Economics*, 139, 1557–1610.

Sætra, H. S., H. Borgebund, and M. Coeckelbergh (2022): “Avoid diluting democracy by algorithms,” *Nature Machine Intelligence*, 4, 804–806.

Saxena, N. A., K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu (2019): “How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106.

——— (2020): “How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations,” *Artificial Intelligence*, 283, 103238.

Schmager, S., C. H. Grøder, E. Parmiggiani, I. Pappas, and P. Vassilakopoulou (2024): “Exploring citizens stances on AI in public services: A social contract perspective,” *Data & Policy*, 6, e19.

Schwardmann, P., E. Tripodi, and J. J. Van der Weele (2022): "Self-persuasion: Evidence from field experiments at international debating competitions," *American Economic Review*, 112, 1118–1146.

Schwartzstein, J. and A. Sunderam (2022): "Shared models in networks, organizations, and groups," Tech. rep., National Bureau of Economic Research.

Scurich, N. and D. A. Krauss (2020): "Publics views of risk assessment algorithms and pretrial decision making." *Psychology, Public Policy, and Law*, 26, 1.

Sidhu, D., B. Magistro, B. A. Stevens, and P. J. Loewen (2024): "Why do citizens support algorithmic government?" *Journal of Public Policy*, 1–19.

Simmons, R. (2017): "Big data and procedural justice: Legitimizing algorithms in the criminal justice system," *Ohio St. J. Crim. L.*, 15, 573.

Simon, J., P. H. Wong, and G. Rieder (2020): "Algorithmic bias and the Value Sensitive Design approach," *Internet policy review*, 9, 1–16.

Simonsohn, U., N. Karlsson, G. Loewenstein, and D. Ariely (2008): "The tree of experience in the forest of information: Overweighing experienced relative to observed information," *Games and Economic Behavior*, 62, 263–286.

Stantcheva, S. (2023): "How to run surveys: A guide to creating your own identifying variation and revealing the invisible," *Annual Review of Economics*, 15, 205–234.

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

Statistisches Bundesamt (2023): "Fast ein Drittel der Erwachsenen in Deutschland ist im Ruhestand," Accessed: 2025-06-06.

Stevenson, M. T. and J. L. Doleac (2024): "Algorithmic risk assessment in the hands of humans," *American Economic Journal: Economic Policy*, 16, 382–414.

Stilgoe, J. (2024): "AI has a democracy problem. Citizens assemblies can help." *Science*, 385, eadr6713.

Suchman, M. C. (1995): "Managing legitimacy: Strategic and institutional approaches," *Academy of management review*, 20, 571–610.

Sunshine, J. and T. R. Tyler (2003): "The role of procedural justice and legitimacy in shaping public support for policing," *Law & society review*, 37, 513–547.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.

Tyler, T. R. (2006): *Why people obey the law*, Princeton university press.

Tyler, T. R. and Y. J. Huo (2002): *Trust in the law: Encouraging public cooperation with the police and courts*, Russell Sage Foundation.

van Deursen, A. J., E. van Laar, E. Helsper, and L. Schneider (2023): "The youth Digital Skills Performance Test Results: Report on the results of real-life information navigation and processing, communication and interaction, and content creation and production skills tasks," .

Venkatesh, V., M. G. Morris, and P. L. Ackerman (2000): "A longitudinal field investigation of gender differences in individual technology adoption decision-making processes," *Organizational behavior and human decision processes*, 83, 33–60.

Waechter, N., E. J. Helsper, L. Schneider, A. J. A. M. van Deursen, and E. van Laar (2021): "Youth Digital Skills Indicator: German Questionnaire. Developed for the ySKILLS Project," `https://yskills.eu` and `https://www.lse.ac.uk/media-and-communications/research/research-projects/disto/surveys`, funded by the European Unions Horizon 2020 Research and Innovation Programme, Grant Agreement No. 870612. Accessed: 2025-06-04.

WANG, C., K. WANG, A. BIAN, R. ISLAM, K. N. KEYA, J. FOULDS, AND S. PAN (2022): "Do humans prefer debiased AI algorithms? A case study in career recommendation," in *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 134–147.

WANG, S. I. AND C. D. MANNING (2012): "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90–94.

WEIZSÄCKER, G. (2010): "Do we follow others when we should? A simple test of rational expectations," *American Economic Review*, 100, 2340–2360.

WISWALL, M. AND B. ZAFAR (2015): "Determinants of college major choice: Identification using an information experiment," *The Review of Economic Studies*, 82, 791–824.

YSKILLS EU PROJECT (2024): "Youth Skills (ySKILLS)," `https://yskills.eu/`, accessed: 2025-06-04.

ZHANG, B. AND A. DAFOE (2019): "Artificial intelligence: American attitudes and trends," *Available at SSRN 3312874.*

——— (2020): "US public opinion on the governance of artificial intelligence," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 187–193.

Anhang der Dissertation

Zusammenfassungen

Chapter 2: *Fragile AI Optimism*

We study how public attitudes toward AI form and shift using an online deliberation experiment with 2,358 UK citizens in the context of criminal justice. First, we replicate prior survey evidence suggesting public support for adopting AI as a decision-support tool, particularly when certain fairness features are met. We then show that this stated support is fragile: it declines significantly with group deliberation, as supporters are 2.6 times more likely than opponents to change their attitudes. Quantitative text analysis indicates that opponents contribute more arguments in group deliberation, both in terms of frequency and topic range, and supporters are more responsive to counterarguments. These results suggest that stated support for AI reflects lower attitude strength as it appears to be easily raised through informational framing but quickly reversed through deliberation. More broadly, they caution against inferring public legitimacy of increased AI deployment from stated support alone.

Wir untersuchen, wie sich öffentliche Einstellungen zu Künstlicher Intelligenz (KI) formen und verändern, und zwar mithilfe eines Online-Deliberationsexperiments mit 2,358 Bürger: im Vereinigten Königreich im Kontext der Strafjustiz. Zunächst replizieren wir frühere Umfrageergebnisse, die auf eine öffentliche Unterstützung für den Einsatz von KI als Entscheidungsunterstützungstool hinweisen insbesondere dann, wenn bestimmte Fairnesskriterien erfüllt sind. Anschließend zeigen wir, dass diese erklärte Unterstützung fragil ist: Sie nimmt durch Gruppendiskussionen deutlich ab, wobei Befürworter:2,6-mal häufiger als Gegner:ihre Meinung ändern. Eine quantitative Textanalyse zeigt, dass Gegner:in Gruppendiskussionen mehr Argumente einbringen sowohl in Bezug auf Häufigkeit als auch thematische Breite und dass Befürworter:stärker auf Gegenargumente reagieren. Diese Ergebnisse deuten darauf hin, dass die erklärte Unterstützung für KI eine geringe Einstellungsstärke widerspiegelt: Sie lässt sich leicht durch informationelle Rahmung erhöhen, kehrt sich jedoch durch Deliberation schnell wieder um. Insgesamt warnen die Ergebnisse davor, aus bloßer Zustimmung auf eine breite öffentliche Legitimität für den vermehrten Einsatz von KI zu schließen.

Chapter 3: *Algorithmic Fairness and Human Discrimination*

Fairness constraints in algorithm design aim to reduce discrimination. Their impact, however, also depends on the adoption of the algorithm by human-decision makers as they typically retain full authority in high-stakes contexts. In a hiring experiment, I first find suggestive evidence that protecting group membership in algorithmic predictions leads individuals to be more conservative in updating their beliefs about candidates based on these predictions. I then find a significant increase in discrimination in their hiring of candidates under this algorithm, driven by those who initially believe that group membership predicts performance. Finally, independent of the algorithm features, about 26% of participants make hiring decisions that cannot be explained by beliefs and are likely based on taste. These results suggest that algorithmic fairness features can paradoxically exacerbate human discrimination based on statistical beliefs by hindering adoption and, unsurprisingly, remain orthogonal to taste-based discrimination.

Ich untersuche, wie sogenannte Fairness-Beschränkungen in der Gestaltung von Algorithmen etwa der Ausschluss geschützter Merkmale wie Geschlecht Diskriminierung beeinflussen. Solche Eingriffe zielen darauf ab, Benachteiligung zu verringern. Ihre Wirkung hängt jedoch maßgeblich davon ab, ob und wie menschliche Entscheidungsträger:die algorithmischen Empfehlungen übernehmen, insbesondere in Kontexten mit hoher Tragweite, in denen Menschen die finale Entscheidungshoheit behalten.

In einem Einstellungsexperiment zeige ich zunächst, dass der Schutz von Gruppenmerkmalen in algorithmischen Vorhersagen dazu führt, dass Personen vorsichtiger in der Aktualisierung ihrer Überzeugungen über Kandidat:reagieren. Anschließend finde ich einen signifikanten Anstieg der Diskriminierung bei ihren Einstellungsentscheidungen unter einem solchen Fairness-Algorithmus insbesondere bei jenen, die zuvor glaubten, dass Gruppenmerkmale Leistungsfähigkeit vorhersagen. Schließlich stelle ich fest, dass etwa 26 % der Teilnehmenden Entscheidungen treffen, die nicht durch ihre Überzeugungen erklärbar sind und vermutlich auf präferenzbasierter Diskriminierung beruhen.

Diese Ergebnisse legen nahe, dass Fairness-Eingriffe in Algorithmen unbeabsichtigt menschliche Diskriminierung verstärken können insbesondere wenn sie die wahrgenommene Aussagekraft der algorithmischen Empfehlungen mindern. Gleichzeitig bleiben sie erwartungsgemäß wirkungslos gegenüber präferenzbasierter Diskriminierung.

Chapter 4: *Social Disparities in Digital Skills: Evidence from Germany*

This study documents gender and socioeconomic gaps in digital skills relevant to the labor market, using a representative German household sample. Men and individuals with a higher level of education demonstrate greater proficiency. Both groups also hold more optimistic beliefs about outperforming others, conditional on actual skills. These belief gaps are not driven by overconfidence, but by underconfidence among women and individuals with lower education backgrounds in the upper tail of the skill distribution. Early-life socioeconomic background is not significantly associated with adult digital skills or beliefs.

In dieser Studie untersuche ich Geschlechter- und sozioökonomische Unterschiede in digitalen Kompetenzen, die auf dem Arbeitsmarkt relevant sind. Grundlage ist eine repräsentative Stichprobe deutscher Haushalte. Die Ergebnisse zeigen, dass Männer sowie Personen mit höherem Bildungsniveau im Durchschnitt über ausgeprägtere digitale Fähigkeiten verfügen. Beide Gruppen schätzen auch ihre eigenen Leistungen im Vergleich zu anderen optimistischer ein und zwar unabhängig davon, wie gut sie tatsächlich abschneiden.

Diese Unterschiede in der Selbsteinschätzung lassen sich jedoch nicht durch eine generelle Selbstüberschätzung erklären. Vielmehr zeigt sich, dass Frauen und Personen mit niedrigerem Bildungsniveau selbst dann zurückhaltender in ihrer Selbsteinschätzung sind, wenn sie objektiv zu den leistungsstärkeren gehören. Es handelt sich also um eine Form von Underconfidence in der oberen Leistungsspanne. Interessanterweise spielt der sozioökonomische Hintergrund in der frühen Kindheit keine signifikante Rolle für digitale Fähigkeiten oder Überzeugungen im Erwachsenenalter.

Liste der aus dieser Dissertation hervorgegangenen Veröffentlichungen

-

Selbstdeklaration bei kumulativen Promotionen

Konzeption / Planung: Formulierung des grundlegenden wissenschaftlichen Problems, basierend auf bisher unbeantworteten theoretischen Fragestellungen inklusive der Zusammenfassung der generellen Fragen, die anhand von Analysen oder Experimenten / Untersuchungen beantwortbar sind. Planung der Experimente / Analysen und Formulierung der methodischen Vorgehensweise, inklusive Wahl der Methode und unabhängige methodologische Entwicklung.

Durchführung: Grad der Einbindung in die konkreten Untersuchungen bzw. Analysen.

Manuskripterstellung: Präsentation, Interpretation und Diskussion der erzielten Ergebnisse in Form eines wissenschaftlichen Artikels.

Die Einschätzung des geleisteten Anteils erfolgt mittels Punkteinschätzung von 1-100%.

Für den zweiten Artikel (Chapter 2) liegt die Eigenleistung für

das Konzept / die Planung bei	50%
die Durchführung bei	75%
der Manuskripterstellung bei	100%

Für den dritten vorliegenden Artikel (Chapter 3) liegt die Eigenleistung bei 100%.

Für den vierten vorliegenden Artikel (Chapter 4) liegt die Eigenleistung bei 100%.

Die vorliegende Einschätzung in Prozent über die von mir erbrachte Eigenleistung wurde mit den am Artikel beteiligten Koautoren einvernehmlich abgestimmt.

______________________   ______________________

Ort/Datum   Doktorand/in

Erklärung

Hiermit erkläre ich, Arna Carolin Wömmel, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

____________________ ____________________

Ort/Datum Doktorand/in

Eidesstattliche Versicherung

Ich, Arna Carolin Wömmel, versichere an Eides statt, dass ich die Dissertation mit dem Titel:

"*On Human Factors in Machine Fairness: Essays in Behavioral Economics*"

selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach § 6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 18. Januar 2017 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

____________________ ____________________

Ort/Datum Doktorand/in