MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

Arian Henning
Pascal Langenbach

Discussion Paper
2024/11

# BRIDGING THE HUMAN-AUTOMATION FAIRNESS GAP: HOW PROVIDING REASONS ENHANCES THE PERCEIVED FAIRNESS OF PUBLIC DECISION-MAKING

# Bridging the Human-Automation Fairness Gap: How Providing Reasons Enhances the Perceived Fairness of Public Decision-Making

**Arian Henning / Pascal Langenbach**
Max Planck Institute for Research on Collective Goods

This version May 6, 2024

## Abstract

Automated decision-making in legal contexts is often perceived as less fair than its human counterpart. This human-automation fairness gap poses practical challenges for implementing automated systems in the public sector. Drawing on experimental data from 4,250 participants in three public decision-making scenarios, this study examines how different reasoning models influence the perceived fairness of automated and human decision-making. The results show that providing reasons enhances the perceived fairness of decision-making, regardless of whether decisions are made by humans or machines. Moreover, the study demonstrates that sufficiently individualized reasoning largely mitigates the human-automation fairness gap. The study thus contributes to the understanding of how procedural elements like giving reasons for decisions shape perceptions of automated government and suggests that well-designed reason giving can improve the acceptability of automated decision systems.

## I. Introduction

The integration of algorithms has gained traction in public decision-making (Engstrom et al., 2020). Algorithms have been used in frequently occurring selection and allocation tasks of public administration, such as selecting tax-audit target organizations (Mehdiyev et al., 2021), admitting students to universities (Kearns & Roth, 2021), or distributing refugees within destination countries to maximize the employment rate (Bansak et al., 2018). Algorithmic public decision-making comes with the promise of increased efficiency, equity, and accuracy compared to alternative systems reliant on human judgment (Grove et al., 2000; Kleinberg et al., 2018). Of course, automated governance poses many challenges in terms of individual justice and potential discrimination (Janssen & Kuk, 2016; Lee et al., 2019; Mendes & Mattiuzzo, 2022; Wu, 2023). A further behavioral challenge for the success of automated government arises from people's reactions to algorithmic decision-making: Even visibly superior algorithmic decision-making solutions often encounter skepticism; the so called 'algorithm aversion' (Castelo et al., 2019; Dietvorst et al., 2015; Jussupow et al., 2020). This skepticism is particularly pronounced in legal settings, where algorithmic decisions are frequently perceived as less fair compared to those made by humans (e.g., Chen et al., 2022; Grimmelikhuijsen, 2023; Hermstrüwer & Langenbach, 2023; Wang, 2018). However, people's fairness perceptions of public decision-making procedures are important for effective governance as fair procedures promote legal compliance and cooperation (e.g., Tyler, 2003, 2006; Tyler & Huo, 2002; Tyler & Jackson, 2014).

The fairness gap between human and automated public decision-making poses practical problems: First, automated governance might run into non-compliance and non-cooperation if people perceive its procedures as unfair, even if automation produces more accurate and equitable outcomes. Second, public decision-makers anticipating public distrust of these procedures might refrain from implementing even beneficial algorithmic decision-making tools in the first place (Nagtegaal, 2021; Simmons, 2018). Apart from this consequentialist perspective, there is deontological value in using public decision-making procedures which the public largely perceives as fair (Juijn et al., 2023; Scurich & Krauss, 2020).

In this study, we explore how accompanying automated administrative decisions with different forms of reasons changes people's fairness perceptions and, in particular, whether this reduces the human-automation fairness gap in public decision-making. We thus contribute to an emerging literature that studies the effects of conventional elements of legal procedures on the perceived fairness of automated decision-making (Chen et al., 2022). Increasing human oversight of automated public decision-making can increase perceived fairness (Hermstrüwer & Langenbach, 2023), as can the implementation of hearing rights (Chen et al., 2022) and providing more comprehensive details about the decision process (Grgić-Hlača et al., 2018; Grimmelikhuijsen, 2023; Kizilcec, 2016; Lee et al., 2019).

Algorithm aversion, however, does not occur universally. Several studies in the context of legal and public decision-making show that, under specific circumstances, automated decision-making is even preferred over purely human decision-making, e.g., in the area of traffic control (Miller & Keiser, 2021), university admissions (Marcinkowski et al., 2020), and law enforcement (Araujo et al., 2020). The perceived fairness of automated decision-making procedures seems to depend on the context (Starke et al., 2022), and so might the effectiveness of potentially fairness increasing procedural features. We therefore study reasoning in three different areas of public decision-making: the reallocation of refugees across the country, the allocation of child daycare places, and university admissions.

Using data from an online vignette experiment with 4,250 participants, we first replicate the overall human-automation fairness gap in the three different contexts of public decision-making. Second, we show that giving sufficiently individualized reasons increases the perceived fairness of (automated) public decision-making; not only compared to procedures without any reasons, but also compared to more formalized, abstract reasons. Third, we find that individualized reasons have a stronger effect on the perceived fairness of automated decision-making than on human decision-making, which narrows, and in some cases effectively closes, the human-automation fairness gap.

## Algorithm Aversion and Algorithmic Fairness in Public Decision-Making

The term 'algorithm aversion' describes the tendency to favor human decision processes over algorithmic ones, even in the presence of evidence for a superior or at least comparable performance of algorithms (Dietvorst et al., 2015; Jago, 2019; Palmeira & Spassova, 2015). More generally speaking, algorithm aversion can be understood as a '*biased assessment of an algorithm which manifests in negative behaviours and attitudes towards the algorithm compared to a human agent'* (Jussupow et al., 2020a, p.4). However, the concept of algorithm aversion remains a subject of study with varying interpretations and partly ambiguous empirical results (cf. Castelo et al., 2019; Jussupow et al., 2020a). This pertains not only to the magnitude of the effect, but also to whether algorithm aversion manifests itself at all or whether, in specific circumstances, the contrasting phenomenon of algorithm appreciation prevails (Hou & Jung, 2021; Logg et al., 2019; You et al., 2022).

Algorithm aversion is differently assessed in empirical studies. A common approach involves the direct comparison between human and machine agents, offering participants the opportunity to select their preferred decision source (see, e.g., Bigman & Gray, 2018). When participants prefer human decisions over algorithms despite similar or better algorithmic performance, this indicates a bias against automated processes (cf. Longoni et al., 2019). Alternatively, studies examine the degree to which individuals consider the judgments of agents by measuring the weight of advice from a human compared to an algorithmic agent in a prediction task (Önkal et al., 2009; You et al., 2022). Finally, survey experiments inquire into the perceived appropriateness, trust, and fairness of decision-making by either human or algorithmic agents (Chen et al., 2022; Diab et al., 2011; Dodge et al., 2019; Hermstrüwer & Langenbach, 2023; Kennedy et al., 2022; Kizilcec, 2016; Madhavan & Wiegmann, 2007; Marcinkowski et al., 2020). If the algorithm performs worse in the respective ratings, this can be interpreted as a sign of algorithm aversion (Bigman & Gray, 2018; Castelo et al., 2019; Jussupow et al., 2020; Madhavan & Wiegmann, 2007).

Algorithmic fairness typically implies that the outcomes generated by an algorithm should be free from discriminatory or unequal impacts (Kilbertus et al., 2017; Shin & Park, 2019; Wachter et al., 2018). This can be defined through mathematical frameworks, such as employing statistical or similarity-based metrics (Dwork et al., 2012; Gajane & Pechenizkiy, 2018; Kusner et al., 2018) to ensure 'tolerable discrimination levels' (Starke et al., 2022).

Alternatively, algorithmic fairness can be founded on the notion that fairness lies in the eye of the addressee (Starke et al., 2022). Fairness is thus a psychological concept and shaped by people's subjective assessments (Binns et al., 2018; Grgić-Hlača et al., 2018; Marcinkowski et al., 2020). Different factors can play a role in how people form their fairness perceptions. Besides the decision outcome itself, people care about decision-making procedures (Lind & Tyler, 1988; Thibaut & Walker, 1975; Tyler, 2006). An expanding literature explores fairness perceptions in automated decision-making within legal frameworks (e.g., Binns et al., 2018; Chen et al., 2022; Grgić-Hlača et al., 2018; Grimmelikhuijsen, 2023; Hermstrüwer & Langenbach, 2023; Marcinkowski et al., 2020).

In our experiment, we follow the second approach, investigating fairness perceptions of hypothetical scenarios in a representative sample of the German population. In line with a large part of the literature studying the relative fairness of human and automated decision-making, overall, we expect to replicate the human-automation fairness gap in our study. This means that – ceteris paribus – human decision-making will receive higher fairness ratings than automated decision-making (Hypothesis 1).

## Reasoning Models

In many settings, public authorities have to give (written) reasons for their decisions. Concerning automated decision-making, there is a broad range of different reasoning models – ranging from local to global, post-hoc to intrinsic explanations, varying in interactivity and quantitative emphasis. In this study, we investigate the perceived fairness of four different reasoning models in (automated) public decision-making. These models can be categorized based on the degree of individualization and information they provide. More abstract and formal explanations might reference decision standards, abstract rules, and guiding principles without explicit subsumption of the individual case, while more individualized reasoning models explain how the specific case at hand has been subsumed under a general decision standard. For legally relevant decision-making that affects fundamental rights using high-risk AI systems, Article 68c of the EU AI-Act, for example, gives the persons affected the right in principle to request a “meaningful explanation on the role of the AI system in the decision-making procedure and the main elements of the decision taken”.[1] Yet, how this explanation should be provided is not determined.

Within individualized reasoning models, one can distinguish causal and counterfactual explanations. Causal explanations aim to pinpoint the factors that directly precipitated an event. This approach is often deterministic, characterizing specific conditions or actions as invariably leading to a certain outcome (Pearl, 2009). Contrastingly, counterfactual explanations use hypotheticals and state how a set of variables would have had to be different for an alternative outcome to be realized (Chou et al., 2022; Pearl, 2013; Wachter et al., 2018; Warren et al., 2023). Human explanations usually strive for causality (Keil, 2006). Causal explanations can thus be regarded as ‘everyday explanations’ (Warren et al., 2023) and resonate with legal justification requirements. Causal and counterfactual explanations are intrinsically interdependent (Mittelstadt et al., 2019; Pearl, 2013; Pearl & Mackenzie, 2018), as causal understanding psychologically and conceptually presupposes the notion of counterfactual thinking (Angrist & Pischke, 2008; Gerstenberg et al., 2021; McCloy & Byrne, 2002; Warren et al., 2023).

For automated decision-making, counterfactuals offer practical advantages: A commonly referenced rationale for employing counterfactual explanations in automated systems is the limitations of machine-learning algorithms in performing causal analysis (Pearl, 2013; Pearl & Mackenzie, 2018). These models technically operate based on correlations and at least initially lack the capability for causal evaluation of different decision variables. Furthermore, more advanced machine-learning algorithms incorporate a multitude of layers and feedback loops, relying on an extensive amount of data. Due to this deep learning architecture, certain models reach a level of opacity that makes a (causal) explanation of which inputs led to which output practically impossible (Murdoch et al., 2019).

While there has been some prior empirical research on various explanation types in public automated decision-making, the behavioral effects of different reasoning models are under-researched. An older literature on expert-systems – not limited to the domain of public decision-making – reports that explanations have a positive effect on attitudes (Clancey, 1983; Neches et al., 1985; Swartout, 1983). Closely related to our experiment, studying the automation of street-level bureaucratic decision-making, Grimmelikhuijsen (2023) finds that providing a

[1] Additionally, Article 13 ensures an “appropriate degree and type of transparency” […] “to enable users to interpret the system’s output and use it appropriately”.

causal explanation for an automated decision can increase the trustworthiness of the decision-making system compared to a setting without any explanation. The fairness effects of counterfactual explanations have also been subject to experimental inquiry: Binns et al. (2019) and Dodge et al. (2019) have put presumably legally valid explanation styles to test, including counterfactual explanations. For the direct comparison of causal and counterfactual explanations, Warren et al. (2023) find in the context of legal driving limits that counterfactual explanations yield higher trust scores than causal explanations. Notwithstanding this limited evidence, the question how counterfactual explanations affect the perceived fairness of public decision-making, also relative to other reasoning models, is far from settled (Wachter et al., 2018).

Regarding the different reasoning models, our study is not limited to automated decisions, but also contributes to the discourse on which reasoning models might enhance perceptions of public decisions per se, encompassing the two different decision modes, automated and human. We therefore address whether counterfactual explanations, often seen as substitutes for causal explanations in complex machine-learning contexts, can also serve as alternatives for traditional explanations of human decisions in terms of their perceived fairness. Consequently, our study does not implement a counterfactual explanation for an actual AI decision-making tool. Instead, in order to allow for comparisons between human and automated as well as causal and counterfactual decisions, we apply stylized representations of the different reasoning models, which are applicable to both human and automated decision-making.

Concerning the fairness effects of the different reasoning models, we expect that more individualized reasoning models, such as causal and counterfactual explanations, will increase fairness perceptions compared to a control treatment without any explanation attached to the decision. This effect is likely to occur primarily due to the higher information density in individualized reasoning models, and should therefore be present in both human and automated decision-making (Hypothesis 2).

The current state of the literature does not allow us to develop directed hypotheses on the comparison between the most individualized reasoning models, that is, causal and counterfactual explanations, in our public decision-making settings, as well as on the potentially differential effects of the different reasoning models on human or automated decision-making. Therefore, the analyses of fairness differences between causal and counterfactual explanations, and in particular of the interaction effects between the reasoning models and the decision modes, are largely explorative in our study.

This paper proceeds as follows: Section II describes the design of our study and the experimental procedures. Section III reports our results, which we discuss in Section IV.

## II. Method

In this section, we explain the experimental design, data collection, decision scenarios, reasoning models, decision modes, and measures of our study. The experiment and supplementary data have been preregistered.

### Design

We conducted an online vignette experiment with a 3x5x2 design: the study used a within-subjects design with three decision scenarios and a between-subjects design with five treatments of different reasoning models, each tested under two different decision modes (human or automated). Participants were presented with the vignettes depicting administrative decision scenarios, and were subsequently provided with the government decision, including the outcome and one of the five reasoning approaches in either the human or the automated decision mode.

### Data collection

We collected 4,250 observations, with participants sourced from Bilendi, ensuring representativeness on quotas for age (in the age range of 18 to 69 years), gender, and education, modeled on the German population. Our participants were roughly evenly distributed across experimental conditions (408 to 436 participants for each of the ten between-subjects groups). To assure data quality, we included two attention checks that had to be answered correctly for participants to complete the study.

### Decision Scenarios

We employed three different contexts of public decision-making in which allocations or selections had to be made.[2] All scenarios cover decisions that would typically be made by government agencies in Germany, the jurisdiction where our participants live. In each of the three real-world decision contexts, the potential for automation is already established (Amarasinghe et al., 2023). However, the technical nature of the decision systems in question varies significantly, ranging from deterministic to dynamic algorithms. Central to our experiment is the concept of automation itself, rather than the detailed technical mechanisms of automation. This focus underscores the broader implications of automated processes, irrespectively of their specific implementations. Using a diverse set of decision scenarios helps us to assess the generalizability of our results across different administrative contexts. The following scenarios were presented in a randomized order to our participants:

1. Reallocation of Refugees:

An asylum seeker from Afghanistan has been initially assigned to the Cologne-Bayenthal reception center. He has requested a transfer to the Hamburg-Rahlstedt center to be closer to his sister, awaiting a decision from the authorities.

2. Allocation of Daycare Places:

A mother is applying for a bilingual daycare spot within walking distance of her home for her 3-year-old daughter, who will begin daycare in six months. She awaits a decision from the local government regarding her applications to three nearby public bilingual daycare centers.

3. University Admission:

Following her successful bachelor's degree in business psychology, a student applies for the master's program in psychology. She has submitted her bachelor's degree certificate and other

[2] The complete wording of all three vignettes can be found in Table A1 in Appendix A.

necessary documents as part of the admission process and now awaits a decision from the public university regarding her application.

**Reasoning Models (treatments)**

Informed by both existing administrative practices and insights gleaned from the literature on explainable artificial intelligence, our experiment features five distinct treatments (CONTROL, PLACEBO, ABSTRACT RULE, CAUSAL, COUNTERFACTUAL). By enriching the information density and varying the style of our explanations, we aim to identify key factors that influence perceptions of fairness of automated and human decisions. *Table 1* displays an example for the different treatments in one decision scenario.[3] Roughly, our explanatory models can be divided into three groups:

1. No explanation (CONTROL)

This treatment serves as the baseline condition, providing participants with the decision outcome only, not presenting any additional information or justification.

2. Explanations without individual case assessment (PLACEBO*,* ABSTRACT RULE)

a) PLACEBO

In this treatment, participants receive the decision outcome along with a statement that the decision was made according to the 'applicable regulations'. While this statement could be seen as underlining the decision-maker's claim not to have acted arbitrarily, a matter of course in the application of administrative law, it lacks any meaningful substantive explanation of the decision. This reasoning model can therefore be seen as related to 'empty' justifications using placebic information (cf. Eiband et al., 2019; Langer et al., 1978).

b) ABSTRACT RULE

Participants in this treatment are presented with the decision outcome and an abstract set of decision criteria, offering insights into the decision-making process without providing individualized case details. Thus, we aim to assess the degree to which participants express a preference for the presentation of the specific decision rule itself. Although this treatment contains more information than the PLACEBO treatment, it does not provide an individualized assessment of the case.

3. Explanations with individual case assessment (CAUSAL*,* COUNTERFACTUAL)

At the level of highest information density, we have opted for causal and counterfactual explanatory models. Although related, these models are analytically distinct. Their implementation in our experiment is conceptually based on the causal and counterfactual explanations of Warren et al. (2023).

a) COUNTERFACTUAL

In the COUNTERFACTUAL treatment, participants receive the decision outcome, the abstract decision criteria, and a counterfactual assessment of the case. Counterfactual explanations use

[3] The wording for our treatments in all decision scenarios can be found in Table A2, Appendix A.

hypotheticals to exemplify how a modification of a decision criterion could have led to an alternative outcome. In algorithmic decision-making, by doing this, counterfactual explanations regularly provide guidance on how individuals can modify their behavior to achieve a more desirable decision outcome potentially, without necessarily explaining the inner logic of the algorithm in use (Mittelstadt et al., 2019; Poyiadzi et al., 2020; Wachter et al., 2018; Warren et al., 2023).

b) CAUSAL

In our CAUSAL treatment, participants receive the decision outcome, abstract decision criteria, and a causal assessment of the decision. Causal explanations aim to convey why a particular decision was made in terms of cause-and-effect relationships between decision criteria (Pearl & Mackenzie, 2018; Warren et al., 2023). They primarily revolve around the identification of the specific factors that directly induce the outcome. For instance, in the context of medical diagnosis, a causal explanation would attribute a patient's illness to a particular virus based on empirical evidence of viral presence. This reasoning style is generally in line with the current standard of administrative law justifications in European legal systems (Olsen et al., 2019).

*Table 1*: Treatments in the daycare scenario:

| **CONTROL** | **PLACEBO** | **ABSTRACT RULE** | **COUNTERFACTUAL** | **CAUSAL** |
|---|---|---|---|---|
| Dear Ms. L,<br>Unfortunately, we are unable to offer you a place in a bilingual daycare center for your child. | | | | |
| | The allocation of daycare places is based on the applicable regulations for the allocation of places in daycare facilities. | The allocation of daycare places is based on the capacities of the selected daycare facilities to ensure that your child is cared for close to home and in line with demand on the desired admission date. | | |
| | | | In the present case, no daycare offer could be made to you. If you had considered bilingual daycare facilities further away from your place of residence instead of the selected daycare facilities, you could have been presented with a daycare offer. | In the present case, no daycare offer could be made to you because there is currently no daycare place available in the bilingual daycare facilities you have specified. |

## Decision Modes

Each treatment is implemented in a between-subjects design in two variants: Either the administrative default is represented by a human signature, showing that a human made the decision, or an introductory sentence in the notice explicitly states that the decision was entirely automated and no signature is provided.[4] By testing the different reasoning models not only for

[4] The exact implementations of the two decision modes can be found in Table A3 in Appendix A.

automated systems, but also for human decision-making, our research may also illuminate potential shortcomings in the justification of human decisions (cf. Zerilli et al., 2019).

### Measures

After each vignette, participants were asked to rate the fairness of the respective decision-making, employing two measures each on a 7-point scale. The first measure was designed to assess the participants' perceived fairness of the decision made (outcome fairness). The second measure was centered on the participants' perceptions of the appropriateness of the applicants' treatment during the administrative decision-making process (procedural fairness). In our analyses, we use the average ratings on these two fairness measures to create a composite "Fairness Index".[5] In addition to the fairness measures, we also assessed participants' reports on how understandable the decision in question was.[6] Moreover, we gathered data on participants' experiences with, and knowledge of, the different decision scenarios.[7]

## III. Results

In this section, we begin with exploring whether human and automated decision-making is evaluated differently. Then, we examine the effects of the different reasoning models on fairness ratings. In the next step, we study whether and how the reasoning models differently affect the evaluation of human and automated decision-making. We start by reporting results on the pooled data across the three different decision scenarios. For this, we collapse the responses each participant gave in the three policy scenarios. Additionally, we use multilevel models to account for the dependency of the responses in the different decision scenarios. Finally, we look into context-specific effects and separately report results for the three public decision-making contexts employed in our study.

### Fairness Differences in Human and Automated Public Decision-Making

In this subsection, we examine whether participants in our study perceive human and automated public decision-making as differently fair. Collapsing the fairness ratings over all decision contexts and all five reasoning models, we replicate the human-automation fairness gap regularly reported in the literature. Overall human decisions are perceived as fairer than automated decisions in public decision-making (N=4250, $p < .001$).[8] Moreover, the human-automation fairness gap in the overall fairness rating is present for all reasoning models separately, that is, in all of our five treatments. *Figure 1* shows the average fairness ratings in each treatment for human and automated decision-making. Differences are (marginally) significant in all the treatments (CONTROL, PLACEBO, and ABSTRACT RULE, p<.001, CAUSAL, p = .059, COUNTERFACTUAL, p = .021), which supports our first hypothesis.

[5] Outcome fairness and procedural fairness are highly positively correlated in our sample (r = .88). We provide a summary analysis for the separate fairness measures in Appendix E.

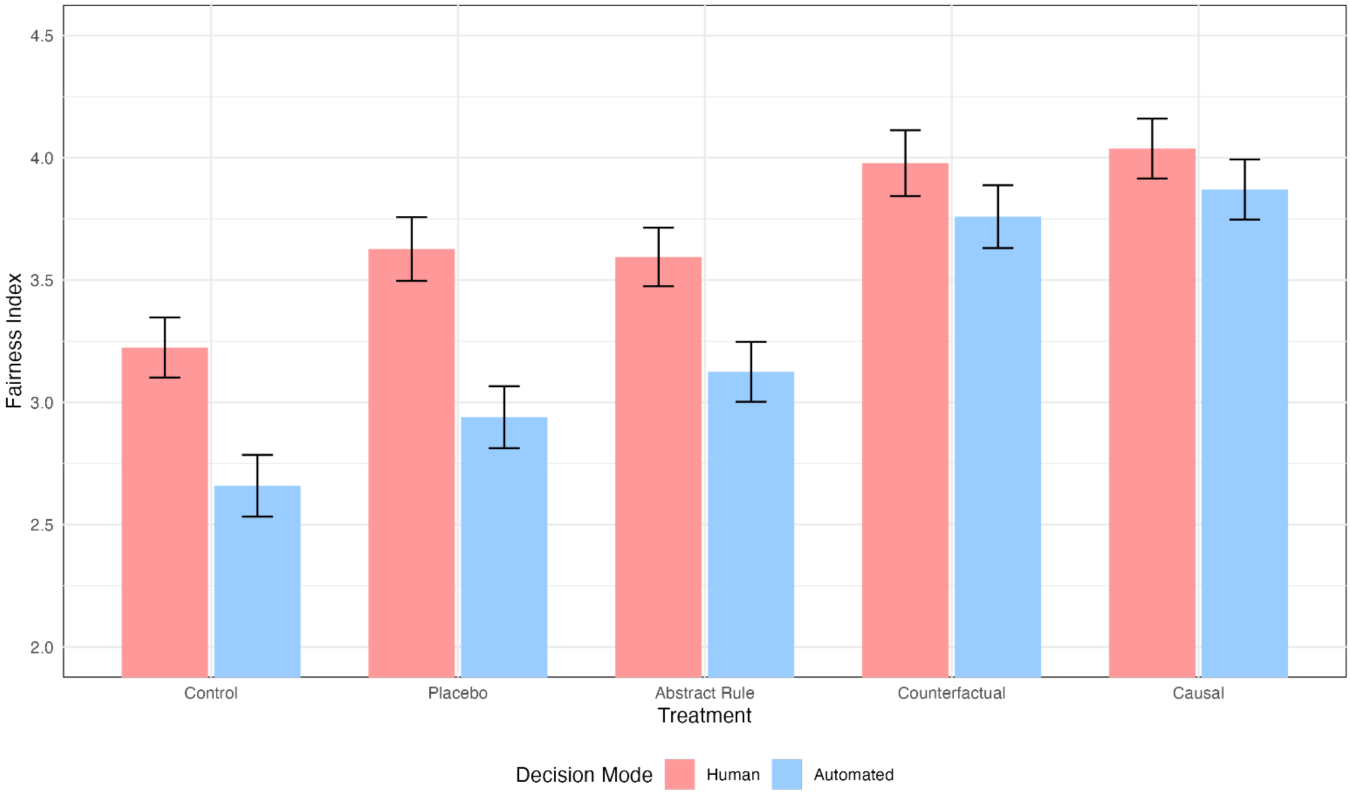
[6] We solely focus on the fairness measure. However, understandability also correlates highly with the fairness measures.

[7] In a post-experimental questionnaire, we also asked for attitudes towards algorithmic decision-making and towards the administration. A list of the questions participants were asked can be found in Appendix B.

[8] We use independent sample t-tests for all group-level comparisons of reasoning model treatments and decision modes. All reported tests are two-sided.

**Result 1:** Overall, in public decision-making, people perceive human decision-making as fairer than automated decision-making.

*Figure 1*: Average fairness ratings for human and automated decision-making under different reasoning models



**Fairness Under Different Reasoning Models**

As is already apparent from *Figure 1*, fairness ratings not only differ between the human and automated decision mode, but also between the different reasoning models. First, we observe that the perceived fairness of human and automated decision-making increases with the degree of information provided. Decision-making without providing any reasons for the decision is perceived as the least fair in human and in automated decision-making (pairwise comparisons of CONTROL vs. each of the other treatments, p < .01).[9] The perceived fairness difference between the CAUSAL/COUNTERFACTUAL treatments and the CONTROL treatment gathers support for Hypothesis 2. Moreover, reasoning models that provide more individualized information, as in the CAUSAL and COUNTERFACTUAL treatments, lead to higher fairness ratings than more abstract explanations in the ABSTRACT RULE treatment or the essentially non-informative explanations in the PLACEBO treatment (comparing CAUSAL | COUNTERFACTUAL vs. ABSTRACT RULE | PLACEBO, p < .001).

**Result 2:** Giving reasons increases the perceived fairness of human and automated public decision-making compared to decision-making without explanations.

[9] All tests in this subsection are conducted separately for human and automated decision-making.

**Result 3:** Reasons with more individualized information lead to higher perceived fairness levels of human and automated decision-making.

Within the more individualized models, different styles of reasoning seem not to affect fairness. While fairness ratings in the COUNTERFACTUAL treatment are descriptively slightly lower than in the CAUSAL treatment under human and automated decision-making, we do not find significant differences between these two treatments (Human: p = .52, Automated: p = .22). People subject to automated decision-making, however, seem to be more sensitive to the subtler differences in the two abstract reasoning models, as they perceive decision-making in the PLACEBO treatment as less fair than in the ABSTRACT RULE treatment (Automated: p = .04, Human: p = .72).

**Result 4:** Causal and counterfactual reasoning models are not perceived as differently fair.

Each participant in our experiment answered fairness questions in three different scenarios of public decision-making presented in a randomized order. Thus, our sample consists of three responses per person for each fairness measure. The reported results have so far been based on the collapsed fairness ratings per person. In the following, we model the dependency in fairness ratings using multilevel models. *Table 2* shows two regression models for the fairness ratings per person in the two different decision modes, human and automated. We control for respondent demographics (age, gender, education, and parenthood) and also include the different scenarios in which subjects made their decisions. The individual-level analyses support the findings, presented earlier, on the differences between the reasoning models. Fairness levels are higher in all treatments with an explanation than in the CONTROL treatment. This holds for both human and automated decision-making. Post-regression Wald tests replicate all further treatment differences reported above (see Table C1 in Appendix C).

*Table 2*: Fairness differences in reasoning models

|  | (1) | (2) |
|---|---|---|
| DV: Fairness | Human | Automated |
| PLACEBO | 0.391*** | 0.297*** |
|  | (0.0891) | (0.0891) |
| ABSTRACT RULE | 0.342*** | 0.480*** |
|  | (0.0896) | (0.0890) |
| CAUSAL | 0.822*** | 1.219*** |
|  | (0.0898) | (0.0896) |
| COUNTERFACTUAL | 0.737*** | 1.100*** |
|  | (0.0906) | (0.0890) |
| Daycare | 0.256*** | 0.126*** |
|  | (0.0429) | (0.0405) |
| University Admission | -0.285*** | -0.263*** |
|  | (0.0429) | (0.0405) |

| | | |
|---|---|---|
| Demographics | ✓ | ✓ |
| Constant | 3.030*** | 2.971*** |
| | (0.253) | (0.270) |
| | | |
| Obs. | 6,363 | 6,387 |
| Groups | 2,121 | 2,129 |

Results from multilevel models. The working sample consists either of observations from the human or the automated decision mode. Observations are grouped at the level of the individual. The reference category for the treatments is the CONTROL treatment, and for the scenarios it is the reallocation-of-refugees scenario. Demographic controls include age, gender, education, and parenthood. Standard errors in parentheses. *** p<0.01

### Closing the Human-Automation Fairness Gap

So far, we have seen an increase in the perceived fairness with more individualized explanations for both human and automated decision-making. However, the individualized reasoning models have another advantage for the employment of automated systems in public decision-making, for they might largely close the fairness gap between human and automated decision-making. As can be seen from *Figure 1*, while fairness in human decision-making dominates fairness in automated decision-making for all reasoning models, the differences between human and automated decision modes become much smaller when more individualized reasons are provided. The fairness advantage of human decision-making in the treatments with individualized reasoning models is less than 50% of the fairness advantage in the other treatments (absolute difference in the fairness ratings between human and automated decision mode in CAUSAL | COUNTERFACTUAL: 0.17 | 0.21; for all other treatments: > .47). This decrease in the fairness differences between human and automated decisions occurs because the CAUSAL and COUNTERFACTUAL reasoning models increase the fairness of automated decision-making more strongly than the fairness of human decision-making. This is supported by the significance of the coefficients for the interactions of the CAUSAL and COUNTERFACTUAL treatment dummies with the dummy for the decision mode in Model 2 of *Table 3*.[10]

**Result 5:** Providing reasons with individualized information can substantially narrow the fairness gap between human and automated decision-making. Individualized reasoning models have a stronger effect on perceived fairness in automated than in human decision-making.

[10] Theoretically, the more pronounced fairness effects of the CAUSAL and COUNTERFACTUAL treatments in the automated decision mode than in the human decision mode could be driven by the fact that participants' response options were restricted from 1 to 7. However, plotting participants' fairness ratings in the two decision modes for each treatment and scenario does not reveal substantial clustering of responses at the upper limit of the scale in any of the treatments or decision modes. This suggests that the scale captured participants' actual fairness perceptions and that the stronger fairness effect of the more individualized reasoning models in the automated decision mode is not a mere artefact of the elicitation method. The cumulative distribution functions for human and automated decision-making in the different treatments can be found in *Figures D1-3* in Appendix D.

*Table 3*: Interaction of decision mode and reasoning model

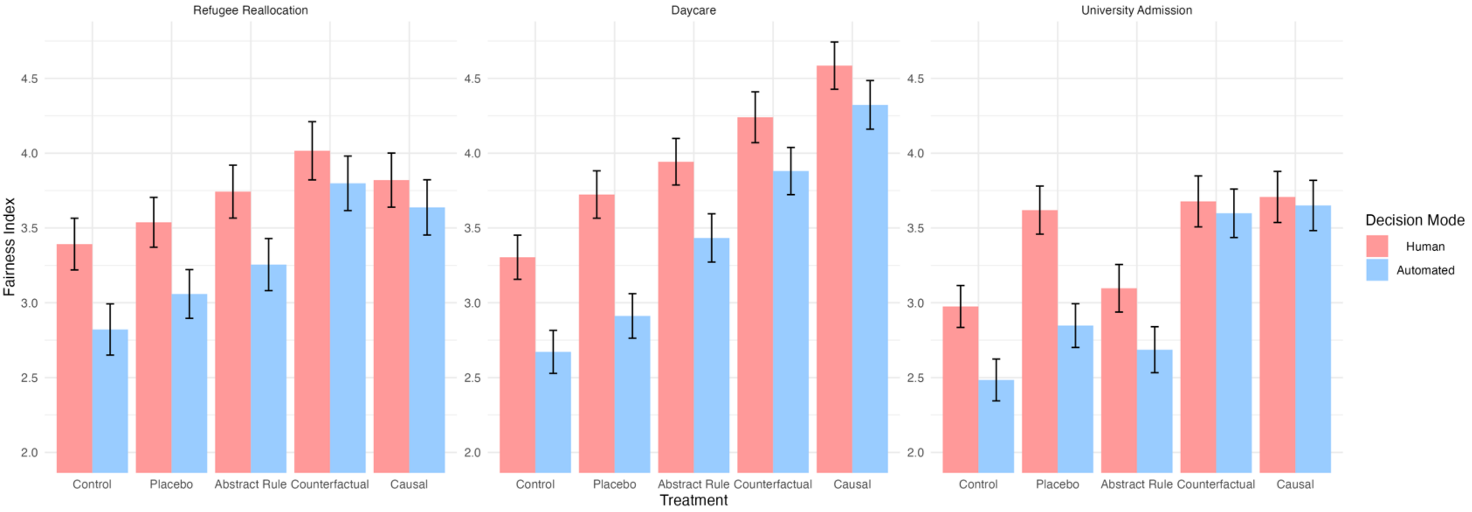
| DV: Fairness | (1) | (2) |
|---|---|---|
| PLACEBO | 0.350*** | 0.398*** |
| | (0.0632) | (0.0891) |
| ABSTRACT RULE | 0.411*** | 0.343*** |
| | (0.0634) | (0.0895) |
| CAUSAL | 1.017*** | 0.822*** |
| | (0.0637) | (0.0898) |
| COUNTERFACTUAL | 0.925*** | 0.739*** |
| | (0.0637) | (0.0905) |
| Automated | -0.421*** | -0.577*** |
| | (0.0403) | (0.0892) |
| PLACEBO*Automated | | -0.0996 |
| | | (0.126) |
| ABSTRACT RULE*Automated | | 0.134 |
| | | (0.126) |
| CAUSAL*Automated | | 0.391*** |
| | | (0.127) |
| COUNTERFACTUAL*Automated | | 0.365*** |
| | | (0.127) |
| Daycare | 0.191*** | 0.191*** |
| | (0.0295) | (0.0295) |
| University Admission | -0.274*** | -0.274*** |
| | (0.0295) | (0.0295) |
| Demographics | ✓ | ✓ |
| Constant | 3.215*** | 3.284*** |
| | (0.187) | (0.190) |
| Obs. | 12,750 | 12,750 |
| Groups | 4,250 | 4,250 |

Results from multilevel models run on the full sample of observations. Observations are grouped at the level of the individual. The reference category for the treatments is the CONTROL treatment, and for the scenarios it is the reallocation-of-refugees scenario. Demographic controls include age, gender, education, and parenthood. Standard errors in parentheses. *** p<0.01

**Context-Specific Effects and Demographics**

Previously, we reported results on the pooled fairness ratings from all three decision scenarios. The different scenarios included the reallocation of refugees, the allocation of daycare places by local government, and university admissions. Fairness ratings for each scenario separately are displayed in *Figure 2*.

*Figure 2*: Fairness Ratings in the Different Decision Scenarios



*Reasoning Models*

Overall, fairness patterns for the different reasoning models look very similar to the aggregated results presented above. However, we also find context-specific treatment effects in our data.[11]

*Refugee Reallocation* – In the refugee scenario, while decision-making with giving reasons is perceived overall as fairer than decision-making without giving reasons (CONTROL vs. all else, except PLACEBO, p < .01), the PLACEBO treatment is not perceived as significantly fairer than the CONTROL treatment in the human decision mode (p = .234). Yet, in the automated decision mode, it is (p = .049). In the human decision mode, the ABSTRACT RULE treatment is not perceived as significantly less fair than the CAUSAL treatment, which provides more individual information (p = .55) while it is in the automated decision mode (p = .003). In this scenario, in both decision modes, the counterfactual reasoning-model is descriptively rated as fairer than all other models. Statistically, the fairness rating of causal and counterfactual explanations is not differently though (Human: p = .147; Automated: p = .222).

*Daycare Place* – In the daycare scenario, decision-making accompanied with reasons is generally perceived as fairer than decision-making without giving reasons (CONTROL vs. all else, p < .03). Also, the treatments with the individualized reasoning models yield higher fairness scores than the treatments in which only abstract reasons are provided (p < .02). In contrast to the aggregated results, however, the ABSTRACT RULE treatment is marginally significantly rated as slightly fairer than the PLACEBO treatment also in the human decision mode (p = .053), and, more importantly, the differences between the CAUSAL and the COUNTERFACTUAL treatment turn out significant in both decision modes in the daycare scenario (Human: p = .004, Automated: p < .001).

[11] Again, group-level results are based on independent sample t-tests.

*University Admission* – Decision-making with reasons is mostly perceived as fairer than decision-making without reasons (p < .001).[12] However, in this scenario, the difference between the CONTROL treatment and the ABSTRACT RULE treatment is only marginally significant in the automated decision mode (p = .056) and not statistically significant in the human decision-mode (p = .260). Moreover, under human decision-making, the PLACEBO explanation leads to a higher fairness score than the ABSTRACT RULE treatment (p < .001). In the automated decision mode, there is no statistical difference between the two treatments (p = .135). The treatments with the individualized-reasoning models have higher fairness scores than the more abstract reasoning models in the automated decision mode and the ABSTRACT RULE treatment in the human decision mode (p < .001). Yet, the treatments with individualized reasons are not perceived as fairer than the PLACEBO treatment when the decision is taken by a human (p > .46). The causal and the counterfactual reasoning models do not receive different fairness ratings either (p > .66).

*Fairness Gap*

We find the human-automation fairness gap in the CONTROL group (diff > .49, p < .001) as well as in the abstract reasoning model-treatments in all the different scenarios (diff >.41, p < .001). However, in the refugee scenario and the university-admission scenario, there are no statistical differences between human and automated decision-making when individualized information is provided, effectively closing the human-automation fairness gap. In the university-admission scenario, the difference is even descriptively neglectable (CAUSAL: diff = .06, p = .64; COUNTERFACTUAL: diff = .08, p = .507). In the refugee-reallocation scenario, the differences remain descriptively more pronounced and approach the marginal significance threshold (CAUSAL: diff = 0.18, p = .166; COUNTERFACTUAL: diff = 0.22, p = .109). Yet, also in the daycare scenario, in which a visible fairness gap between human and automated decision-making remains in the individualized treatments (CAUSAL: diff = .26, p = .023; COUNTERFACTUAL: diff = .36, p = .002), these differences are considerably smaller than in all other treatments (CONTROL: diff = .63; PLACEBO: diff = .81; ABSTRACT RULE: diff = .51).

*Demographics*

In *Table 4*, we report results from ordinary least squares regression estimations run for each decision scenario separately. We include demographic variables as well as assessments of participants' knowledge and experience with the respective domains in the models. We generally find that higher age is correlated with higher fairness ratings, whereas women rate all decisions as less fair. Not surprisingly, people with children like the decision-making with the negative admission outcome in the daycare scenario less than people who do not have children. Overall domain knowledge and experience lead to higher fairness ratings across all decision contexts.

[12] Smallest p-value for tests of the CONTROL treatment against all other treatments except the ABSTRACT RULE treatment.

*Table 4*: Demographics

| DV: Fairness | (1)<br>Refugee<br>Reallocation | (2)<br>Daycare | (3)<br>University<br>Admission |
|---|---|---|---|
| Placebo | 0.210** | 0.336*** | 0.510*** |
| | (0.0878) | (0.0788) | (0.0786) |
| Abstract Rule | 0.400*** | 0.693*** | 0.166** |
| | (0.0880) | (0.0790) | (0.0788) |
| Causal | 0.629*** | 1.461*** | 0.957*** |
| | (0.0884) | (0.0794) | (0.0792) |
| Counterfactual | 0.800*** | 1.061*** | 0.904*** |
| | (0.0885) | (0.0794) | (0.0793) |
| Automated | -0.381*** | -0.513*** | -0.360*** |
| | (0.0559) | (0.0502) | (0.0501) |
| Age | 0.004* | 0.004** | -0.003 |
| | (0.0021) | (0.0020) | (0.0019) |
| Gender (f) | -0.311*** | -0.172*** | -0.290*** |
| | (0.0571) | (0.0510) | (0.0509) |
| Education | ✓ | ✓ | ✓ |
| Parenthood | 0.069 | -0.153*** | -0.030 |
| | (0.0605) | (0.0586) | (0.0540) |
| Domain Knowledge | 0.186*** | 0.080*** | 0.142*** |
| | (0.0185) | (0.0160) | (0.0176) |
| Domain Experience | 0.247*** | 0.214*** | 0.256*** |
| | (0.0872) | (0.0607) | (0.0714) |
| Constant | 2.254*** | 2.345*** | 2.306*** |
| | (0.3175) | (0.2709) | (0.2840) |
| Observations | 4,250 | 4,250 | 4,250 |
| R-squared | 0.070 | 0.123 | 0.091 |

Results from OLS models. The reference category for the treatments is the CONTROL treatment. The gender dummy equals 1 if the participant reported to be female, and the parenthood dummy equals 1 if the participant indicated that they had children. Domain knowledge and domain experience were answered on a scale of 1 to 7, on which participants indicated whether they had experience with or knowledge about the procedures in question. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

## IV. Discussion and Conclusion

Efficiency gains from automated decision-making in public administration must always be balanced against the disadvantages of their implementation. One possible disadvantage is the potentially reduced compliance due to a biased assessment of algorithmic decision-making

(Dietvorst et al., 2015; Jussupow et al., 2020). As procedural-justice research has consistently shown, the perceived fairness of public decision-making affects legal compliance and the acceptance of decisions (Tyler, 2003, 2006; Tyler & Huo, 2002). Fairness perceptions of automated decision systems can vary based on several factors such as their performance (Yeomans et al., 2019), their degree of autonomy (Hermstrüwer & Langenbach, 2023; Komiak & Benbasat, 2006; Nissen & Sengupta, 2006), and the expertise exhibited by human decision-makers (Önkal et al., 2009). From a policy perspective, the fairness gap between human and automated decision-making has frequently been met with a call for increased transparency (cf. Olsen et al., 2019). Various regions, most prominently the European Union, subject automated decision-making systems to transparency regulations,[13] primarily based on the idea of disclosing the inner logic of the respective algorithms (Almeida et al., 2022; Busuioc et al., 2023; Esposito, 2022a; Gryz & Rojszczak, 2021). Empirical research indicates that providing more technical information is just one of several paths toward refining automated administrative practices (Grimmelikhuijsen, 2023; Kizilcec, 2016). The mere fact that information is provided does not necessarily lead to well-informed recipients of (automated) decisions (Bawden & Robinson, 2020; Ndumu, 2020; Pieters, 2011). Too many, too long, or too complicated explanations have been shown to reduce trust in automated decisions (Kizilcec, 2016). Another approach to enhance trust and acceptance might involve explanations as a communicative process rather than the mere provision of information (Esposito, 2022b, 2022a). Of course, understanding transparency as a communicative act does not mean the release of an accountable development of the technologies used for democratic control (Busuioc et al., 2023).

We have provided experimental evidence on the perceived fairness of human and automated public decision-making under different forms of explanations. The overall finding, namely that human decision-making is perceived as fairer than automated decision-making, is in line with the existing literature that often highlights the human-automation fairness gap (Starke et al., 2022). Our study extends this literature by demonstrating that, while this gap largely persists for decision-making processes under different reasoning models, more individualized reasoning models can considerably reduce it. Notably, in some decision contexts, individualized reasons even make the fairness difference disappear.

Considering the level differences in fairness ratings between the different scenarios, the aforementioned effect of information overload could partially explain the relatively low perceived fairness of the decisions under the more extensive reasoning models in the university-admission scenario (see *Figure 2*). In this case, the decision rule provided was considerably longer than in the other two scenarios. Length and complexity of an explanation might affect its evaluation (Amarasinghe et al., 2023). However, our experimental setup does not allow us to identify the reason for these differences between decision contexts, as other attributes of the different decision scenarios might also have affected participant's fairness evaluations.

Regarding the different reasoning models, we show that providing any kind of reasons for decisions significantly enhances the perceived fairness of both human and automated decision-making. This result reinforces the importance of reason-giving in both human and automated

[13] See, e.g., Art. 15 (1) lit. h, Art. 22 GDPR; for high-risk AI applications, see Art. 13, 68c EU AI Act.

public decision systems. Yet, at least in the context of public administration, some jurisdictions provide exceptions for otherwise legally demanded reason-giving if decision-making is automatized.[14] Even an apparently ‘empty’ explanation proves to be more effective than providing no explanation whatsoever. Referring to the fact that decision-making was rule-bound, or providing insights into the abstract decision criteria, could be perceived as a minimum standard of explanation. Our results indicate that public agencies can already profit in terms of perceived fairness if they only use rather reduced reasoning models which are generally easily implementable in automated decision-making.

The reported differences in fairness perceptions across reasoning models provide further insights for the design of automated decision-making systems. Reasoning models that offer individualized explanations lead to higher fairness ratings than those providing more abstract explanations or non-informative explanations. These individualized explanations demonstrate a stronger impact on fairness perceptions in automated decision-making compared to human decision-making. This implies that the development of automated systems that integrate detailed, context-specific reasoning can enhance their acceptability by the public, potentially mitigating some of the biases inherent in the reception of automated decision-making. From a policy perspective, developing accessible explanatory methods for automated decision-making might be helpful.

Of course, deciding on the optimal reasoning model in automated governance not only, and maybe not even mainly, depends on a model’s effect on fairness perceptions and compliance or cooperation rates. Reasoning models also have to be technical feasible, legally valid, and practically useful. Finding the optimal reasoning model is therefore an inherently interdisciplinary task. The more extensive reasoning models become, the more difficult they are to implement technically. This is particularly evident in the case of causal and counterfactual explanations. Unlike counterfactual explanations, causal explanations cannot be adapted by both human and algorithmic agents without objections. Counterfactual explanations are offered as practical alternatives for explaining the results of complex machine-learning algorithms. Nevertheless, multiple counterfactual explanations often exist, and selecting the most appropriate one continues to be a challenge (Sokol & Flach, 2019). The choice of a reasoning model will largely depend on the type and complexity of the decision. For example, counterfactual explanations are supposed to be actionable (Poyiadzi et al., 2020; Wachter et al., 2018; Warren et al., 2023). Yet, this is practically limited if they refer to variables beyond the user’s control (Poyiadzi et al., 2020; Wachter, 2022). The use of more individualized reasoning models can also be legally challenged, for instance if truthful counterfactual explanations involve legally unacceptable decision criteria (Goethals et al., 2023; Wachter, 2022).

One particular goal of our study was to obtain insights into how the analysis of computational explainability models can also be used for human decision-making in legal contexts; and conversely, how reasoning models initially designed for human decision-making fare for automated systems. In order to do this, we designed reasoning models that capture core features of the different explanation styles, but were equally applicable to both human and automated

[14] See, for example, Section 39 (2) of the German Administrative Procedure Act.

decision-making systems. To do this, we had to simplify. For example, our counterfactual explanations used the same decision criteria as were used in the causal explanations. In real-world settings, the criteria presented in causal and counterfactual explanations need not be identical. Moreover, the main applications of counterfactual explanations are automated models that cannot provide causal explanations in a reasonable way. Therefore, even if deciding in the same cases, decision-making between human actors and automated systems might differ, and so might the explanation feasible for each decision mode. A further simplification lies in the fact that we studied decisions with rather clear decision factors which had direct causal paths to the decision outcome. It remains an empirically open question how reasoning models will perform when decision-making is more complex, for example leading to more contested decisions, usually requiring some sort of discretion, or relying on more correlational decision criteria. Apparently, the research on reasoning in human and automated public administration is still in its early stages; it is therefore for future research to explore these questions – and many more.

## V. Acknowledgments

We are grateful for helpful comments on this work by Stefan Bechtold, Alexander Egberts, Christoph Engel, Christian Henning, Yoan Hermstrüwer, Johannes Kruse, Christoph Möllers, Martin Sternberg, and Alexander Stremitzer. We likewise profited from the feedback we received at presentations given at ETH Zurich, Humboldt University of Berlin and the Max Planck Institute for Research on Collective Goods in Bonn. We thank Léon Bartosch for implementing the survey in Qualtrics and Brian Cooper for language help.

## VI. References

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics*, *2*(3), 377–387. https://doi.org/10.1007/s43681-021-00077-w

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, *5*, e5. https://doi.org/10.1017/dap.2023.2

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. https://doi.org/10.1515/9781400829828

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

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*(6373), 325–329.

Bawden, D., & Robinson, L. (2020). *Information Overload: An Overview*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1360

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage" Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14.

Busuioc, M., Curtin, D., & Almada, M. (2023). Reclaiming transparency: Contesting the logics of secrecy within the AI Act. *European Law Open*, *2*(1), 79–105. https://doi.org/10.1017/elo.2022.47

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

Chen, B. M., Stremitzer, A., & Tobia, K. (2022). Having Your Day in Robot Court. *Harvard Journal of Law & Technology*, *36*(1), 128.

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59–83. https://doi.org/10.1016/j.inffus.2021.11.003

Clancey, W. J. (1983). The epistemology of a rule-based expert system—A framework for explanation. *Artificial Intelligence*, *20*(3), 215–251. https://doi.org/10.1016/0004-3702(83)90008-5

Diab, D. L., Pui, S.-Y., Yankelevich, M., & Highhouse, S. (2011). Lay Perceptions of Selection Decision Aids in US and Non-US Samples. *International Journal of Selection and Assessment*, *19*(2), 209–216. https://doi.org/10.1111/j.1468-2389.2011.00548.x

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

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science*

*Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312787

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *NYU School of Law, Public Research Paper No. 20-54*. https://doi.org/10.2139/ssrn.3551505

Esposito, E. (2022a). Does Explainability Require Transparency? *Sociologica*, *16*(3), Article 3. https://doi.org/10.6092/issn.1971-8853/15804

Esposito, E. (2022b). Transparency versus explanation: The role of ambiguity in legal AI. *Journal of Cross-Disciplinary Research in Computational Law*, *1*(2), Article 2. https://journalcrcl.org/crcl/article/view/10

Gajane, P., & Pechenizkiy, M. (2018). *On Formalizing Fairness in Prediction with Machine Learning* (arXiv:1710.03184). arXiv. https://doi.org/10.48550/arXiv.1710.03184

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975. https://doi.org/10.1037/rev0000281

Goethals, S., Martens, D., & Calders, T. (2023). PreCoF: Counterfactual explanations for fairness. *Machine Learning*. https://doi.org/10.1007/s10994-023-06319-8

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Proceedings of the 2018 World Wide Web Conference*, 903–912. https://doi.org/10.1145/3178876.3186138

Grimmelikhuijsen, S. (2023). Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making.

*Public Administration Review*, *83*(2), 241–262. https://doi.org/10.1111/puar.13483

Grove, W., Zald, D., Lebow, B., Snitz, B., & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment*, *12*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gryz, J., & Rojszczak, M. (2021). Black box algorithms and the rights of individuals: No easy solution to the "explainability" problem. *Internet Policy Review*, *10*(2). https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability

Hermstrüwer, Y., & Langenbach, P. (2023). Fair governance with humans and machines. *Psychology, Public Policy, and Law, 29 (4)*, 525-548. https://doi.org/10.1037/law0000381

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 477:1-477:25. https://doi.org/10.1145/3479864

Jago, A. S. (2019). Algorithms and Authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

Juijn, G., Stoimenova, N., Reis, J., & Nguyen, D. (2023). Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 775–785. https://doi.org/10.1145/3600211.3604677

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Publications of*

*Darmstadt Technical University, Institute for Business Studies (BWL)*, Article 138565. https://ideas.repec.org//p/dar/wpaper/138565.html

Kearns, M., & Roth, A. (2021). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. *Perspectives on Science and Christian Faith*, *73*(1), 55–56. https://doi.org/10.56315/PSCF3-21Kearns

Keil, F. C. (2006). Explanation and Understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kennedy, R. P., Waggoner, P. D., & Ward, M. M. (2022). Trust in Public Policy Algorithms. *The Journal of Politics*, *84*(2), 1132–1148. https://doi.org/10.1086/716283

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. *Advances in Neural Information Processing Systems*, *30*, 656–666. https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Komiak, S. Y. X., & Benbasat, I. (2006). The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly*, *30*(4), 941–960. https://doi.org/10.2307/25148760

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2018). Counterfactual Fairness. *Advances in Neural Information Processing Systems, 30,* 4069–4079. https://doi.org/10.48550/arXiv.1703.06856

Langer, E. J., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, *36*(6), 635–642. https://doi.org/10.1037/0022-3514.36.6.635

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359284

Lind, E. A., & Tyler, T. R. (1988). *The Social Psychology of Procedural Justice*. Springer Science & Business Media.

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

Longoni, C., Bonezzi, A., & Morewedge, C. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*. https://doi.org/10.1093/jcr/ucz013

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–130. https://doi.org/10.1145/3351095.3372867

McCloy, R., & Byrne, R. M. J. (2002). Semifactual "even if" thinking. *Thinking & Reasoning*, *8*(1), 41–67. https://doi.org/10.1080/13546780143000125

Mehdiyev, N., Houy, C., Gutermuth, O., Mayer, L., & Fettke, P. (2021). Explainable

Artificial Intelligence (XAI) Supporting Public Administration Processes – On the Potential of XAI in Tax Audit Processes. In Ahlemann, F., Schütte, R., Stieglitz, S. (Eds.), *Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation*, *46*, 413–428. https://doi.org/10.1007/978-3-030-86790-4_28

Mendes, L. S., & Mattiuzzo, M. (2022). Algorithms and Discrimination: The Case of Credit Scoring in Brazil. In M. Albers & I. W. Sarlet (Eds.), *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches*, *96*, 407–443. https://doi.org/10.1007/978-3-030-90331-2_17

Miller, S. M., & Keiser, L. R. (2021). Representative Bureaucracy and Attitudes Toward Automated Decision Making. *Journal of Public Administration Research and Theory*, *31*(1), 150–165. https://doi.org/10.1093/jopart/muaa019

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, *38*(1), 101536. https://doi.org/10.1016/j.giq.2020.101536

Ndumu, A. (2020). Toward a new understanding of immigrant information behavior: A survey study on information access and information overload among US Black diasporic immigrants. *Journal of Documentation*, *76*(4), 869–891. https://doi.org/10.1108/JD-04-2019-0066

Neches, R., Swartout, W. R., & Moore, J. D. (1985). Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development. *IEEE Transactions on Software Engineering*, *SE-11*(11), 1337–1351. https://doi.org/10.1109/TSE.1985.231882

Nissen, M. E., & Sengupta, K. (2006). Incorporating Software Agents into Supply Chains: Experimental Investigation with a Procurement Task. *MIS Quarterly*, *30*(1), 145–166. https://doi.org/10.2307/25148721

Olsen, H. P., Slosser, J. L., Hildebrandt, T. T., & Wiesener, C. (2019). What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration. *iCourts Working Paper Series No. 162, 2019, University of Copenhagen Faculty of Law Research Paper No. 2019-84*. https://doi.org/10.2139/ssrn.3402974

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

Palmeira, M., & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing*, *49*, 302–326. https://doi.org/10.1108/EJM-07-2013-0390

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2013). Structural Counterfactuals: A Brief Introduction. *Cognitive Science*, *37*(6), 977–985. https://doi.org/10.1111/cogs.12065

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (1st ed.). Basic Books, Inc.

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53–64. https://doi.org/10.1007/s10676-010-9253-3

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and Actionable Counterfactual Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350. https://doi.org/10.1145/3375627.3375850

Scurich, N., & Krauss, D. A. (2020). Public's views of risk assessment algorithms and pretrial decision making. *Psychology, Public Policy, and Law*, *26*(1), 1–9. https://doi.org/10.1037/law0000219

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Simmons, R. (2018). Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System. *U.C. Davis Law Review*, *52*(2), 1067–1118.

Sokol, K., & Flach, P. (2019). Counterfactual explanations of machine learning predictions: 2019 AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019. *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*, 2301. http://www.scopus.com/inward/record.url?scp=85060588736&partnerID=8YFLogxK

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

Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, *21*(3), 285–325. https://doi.org/10.1016/S0004-3702(83)80014-9

Thibaut, J. W., & Walker, L. (1975). *Procedural Justice: A Psychological Analysis*. Hillsdale, N.J.: L. Erlbaum Associates.

Tyler, T. R. (2003). Procedural Justice, Legitimacy, and the Effective Rule of Law. *Crime and Justice*, *30*, 283–357. https://doi.org/10.1086/652233

Tyler, T. R. (2006). *Why People Obey the Law*. Princeton University Press. https://doi.org/10.2307/j.ctv1j66769

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

Tyler, T. R., & Jackson, J. (2014). Popular legitimacy and the exercise of legal authority: Motivating compliance, cooperation, and engagement. *Psychology, Public Policy, and Law*, *20*(1), 78–95. https://doi.org/10.1037/a0034514

Wachter, S. (2022). The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law. *Tulane Law Review, 97*(2)*,* 149. https://doi.org/10.2139/ssrn.4099100

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 842–887. https://doi.org/10.48550/arXiv.1711.00399

Wang, A. J. (2018). *Procedural Justice and Risk-Assessment Algorithms* (SSRN Scholarly Paper 3170136). https://doi.org/10.2139/ssrn.3170136

Warren, G., Byrne, R. M. J., & Keane, M. T. (2023). Categorical and Continuous Features in Counterfactual Explanations of AI Systems. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 171–187. https://doi.org/10.1145/3581641.3584090

Wu, Y. (2023). Data Governance and Human Rights: An Algorithm Discrimination Literature Review and Bibliometric Analysis. *Journal of Humanities, Arts and Social Science*, *7*(1), 128–154. https://doi.org/10.26855/jhass.2023.01.018

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting

Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, *32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

## Appendix A: Vignettes

Table A1: Decision scenarios

| **Refugee Reallocation** | **Daycare** | **University Admission** |
|---|---|---|
| M fled from Afghanistan to Germany in 2022 and is currently undergoing asylum proceedings.<br>On his arrival in Germany, M was assigned to the Cologne-Bayenthal initial reception center.<br>This decision is binding for M, i.e., M must live in the assigned facility. However, it is possible to apply for reallocation to another facility. As M's sister already lives in Hamburg, he submits an application to the responsible authority for reallocation to the Hamburg-Rahlstedt initial reception center. M states his sister's place of residence in the application. The competent authority then issues the following decision: | L is looking for a daycare place for her 3-year-old daughter Z. As L mainly works from home, she prefers a daycare center within walking distance of her home. She would also like a bilingual daycare center for her child. Six months before Z is due to start at the daycare center, L applies to enroll her daughter in three different bilingual daycare centers close to her home. The competent authority issues the following decision to L: | A applies for a Master's degree in Psychology at the University of Düsseldorf after successfully completing her Bachelor's degree in Business Psychology at the University of Cologne (final grade 1.90). As part of the admission procedure, A submits her Bachelor's certificate and all other necessary documents. The University of Düsseldorf issues the following decision to A: |

Table A2: Treatments in the different decision scenarios

| Treatment | **Refugee Reallocation** | **Daycare** | **University Admission** |
|---|---|---|---|
| **Control** | Dear Mr. M,<br>Your application for relocation to the initial reception center in Hamburg-Rahlstedt has been rejected. | Dear Ms. L,<br>Unfortunately, we are unable to offer you a place in a bilingual daycare center for your child. | Dear Ms. A,<br>Thank you for your interest in studying at the University of Düsseldorf.<br>Unfortunately, you cannot be admitted to the Master's program in Psychology. |
| **Placebo**<br>*(Control + Placebo)* | The rejection of the application is based on the current regulations on the reallocation of asylum seekers. | The allocation of daycare places is based on the applicable regulations for the allocation of places in daycare facilities. | The admission decision is based on the applicable regulations for the allocation of study places. |
| **Abstract Rule**<br>*(Control + Abstract Rule)* | Reallocation is only possible for reasons of family reunification and other reasons of comparable importance. These include, for example, medical/therapeutic reasons and permanent employment. In the context of family reunification, only spouses and minor children are considered. | The allocation of daycare places is based on the capacities of the selected daycare facilities to ensure that your child is cared for close to home and in line with demand on the desired admission date. | Admission to the Master's degree program is based on the admission limit for the Master's degree program in Psychology (final grade: 2.10) and the other requirements in accordance with the admission and admission regulations for the "Master of Science" degree course in Psychology at the University of Düsseldorf. The prerequisite is a relevant degree within the meaning of § 3 of the admission regulations. |

| | | | |
|---|---|---|---|
| **Counterfactual**<br>*(Control + Abstract Rule + Counterfactual case assessment)* | The requirements for reallocation are not met in your case. If your wife or minor child were living in Hamburg instead of your sister, your application would have been granted. | In the present case, no daycare offer could be made to you. If you had considered bilingual daycare facilities further away from your place of residence instead of the selected daycare facilities, you could have been presented with a daycare offer. | You could not be admitted to the Master's degree program in Psychology. If you had earned 10 credit points in physiological and biological psychology instead of 8 credit points in your Bachelor's degree, you would have fulfilled § 3 of the admission regulations and would have been admitted to the Master's degree program. |
| **Causal**<br>*(Control + Abstract Rule + Causal case assessment)* | The conditions for reallocation are not met in your case because none of the family members mentioned live in Hamburg and no other reasons of comparable weight are apparent. | In the present case, no daycare offer could be made to you because there is currently no daycare place available in the bilingual daycare facilities you have specified. | You could not be admitted to the Master's degree program in Psychology, in particular because you did not earn the required number of credit points in physiological and biological psychology as defined in § 3 of the admission regulations during your Bachelor's degree program with 8 credit points. |

Table A3: Decision Modes

| | **Human** | **Automated** |
|---|---|---|
| Remark on decision mode | - | This decision was made completely automatically and without human involvement. |
| Signature | Yours sincerely,<br>Meyer | This letter was generated by machine and is therefore valid without a signature. |

## Appendix B: Questionnaire

Fairness Ratings conducted after each vignette:

*How fairly do you rate the decision made by the authorities?*
Very unfair (1) - Very fair (7)

*How comprehensible do you find the decision made?*
Not comprehensible at all (1) - Very comprehensible (7)

*How appropriately was M treated?*
Very inappropriately (1) - Very appropriately (7)

Attention checks conducted after the refugee reallocation vignette and the university-admission vignette

Refugee reallocation: *In which city does M's sister live?*
Hamburg / Berlin / Munich

University Admission: *What subject did A apply for?*
Law / Chemistry / Psychology

Additional questions asked at the end of the study:

*How familiar are you with how algorithmic or automated decisions work?*
Not at all familiar (1) - Very familiar (7)

*How convinced are you that algorithms can create prejudices or discriminate against certain groups of people?*
Not at all convinced (1) - Very convinced (7)

*To what extent are you prepared to trust the decisions of algorithms in important decision-making situations?*
Not at all willing (1) - Very willing (7)

*How high is your level of trust in the ability of the public administration to address the needs of citizens adequately?*
Very low trust (1) - Very high trust (7)

*If you think back to your own experiences with public authorities, how fair do you think the procedures and processes in public administration are?*
Not fair at all (1) - Very fair (7)

*How familiar are you with the way refugees are distributed in Germany?*
Not at all familiar (1) - Very familiar (7)

*Do you personally, or does someone close to you, have experience with the distribution of refugees in Germany?*
Yes / No

*How familiar are you with the way daycare places are distributed in your place of residence?*

Not familiar at all (1) - Very familiar (7)

*Have you personally, or has someone close to you, had any experience with the distribution of daycare places?*
Yes / No

*How familiar are you with the selection procedures for restricted admission degree programs in Germany?*
Not at all familiar (1) - Very familiar (7)

*Have you personally, or has someone close to you, had any experience with admission to restricted degree programs?*
Yes / No

## Appendix C: Post-Regression Wald Tests for Treatment Differences

Table C1: Post-regression Wald test run after the regression estimations reported in Table 2.

| | | Automated |
|---|---|---|
| PLACEBO vs. ABSTRACT RULE | p = .581 | p = .04 |
| PLACEBO vs. CAUSAL | p < .001 | p < .001 |
| PLACEBO vs. COUNTERFACTUAL | p < .001 | p < .001 |
| ABSTRACT RULE vs. CAUSAL | p <. 001 | p < .001 |
| ABSTRACT RULE vs. COUNTERFACTUAL | p <. 001 | p < .001 |
| CAUSAL vs. COUNTERFACTUAL | p = .347 | p = .186 |

## Appendix D: Empirical Cumulative Distribution Functions

Figure D1: Empirical Cumulative Distribution Functions of Fairness Index by Decision Mode for each Treatment (Refugee Reallocation)

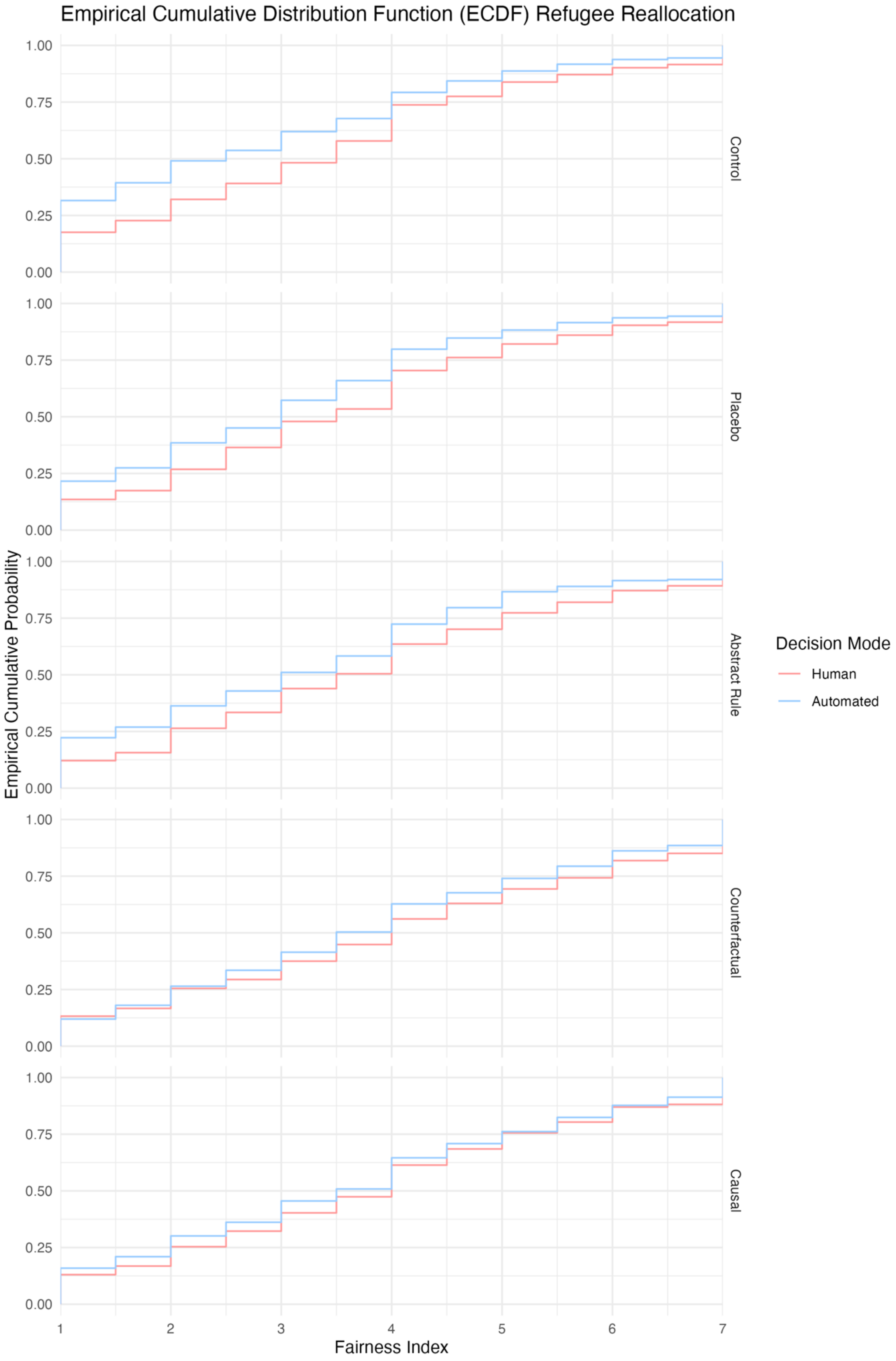


Figure D2: Empirical Cumulative Distribution Functions of Fairness Index by Decision Mode for each Treatment (Daycare)

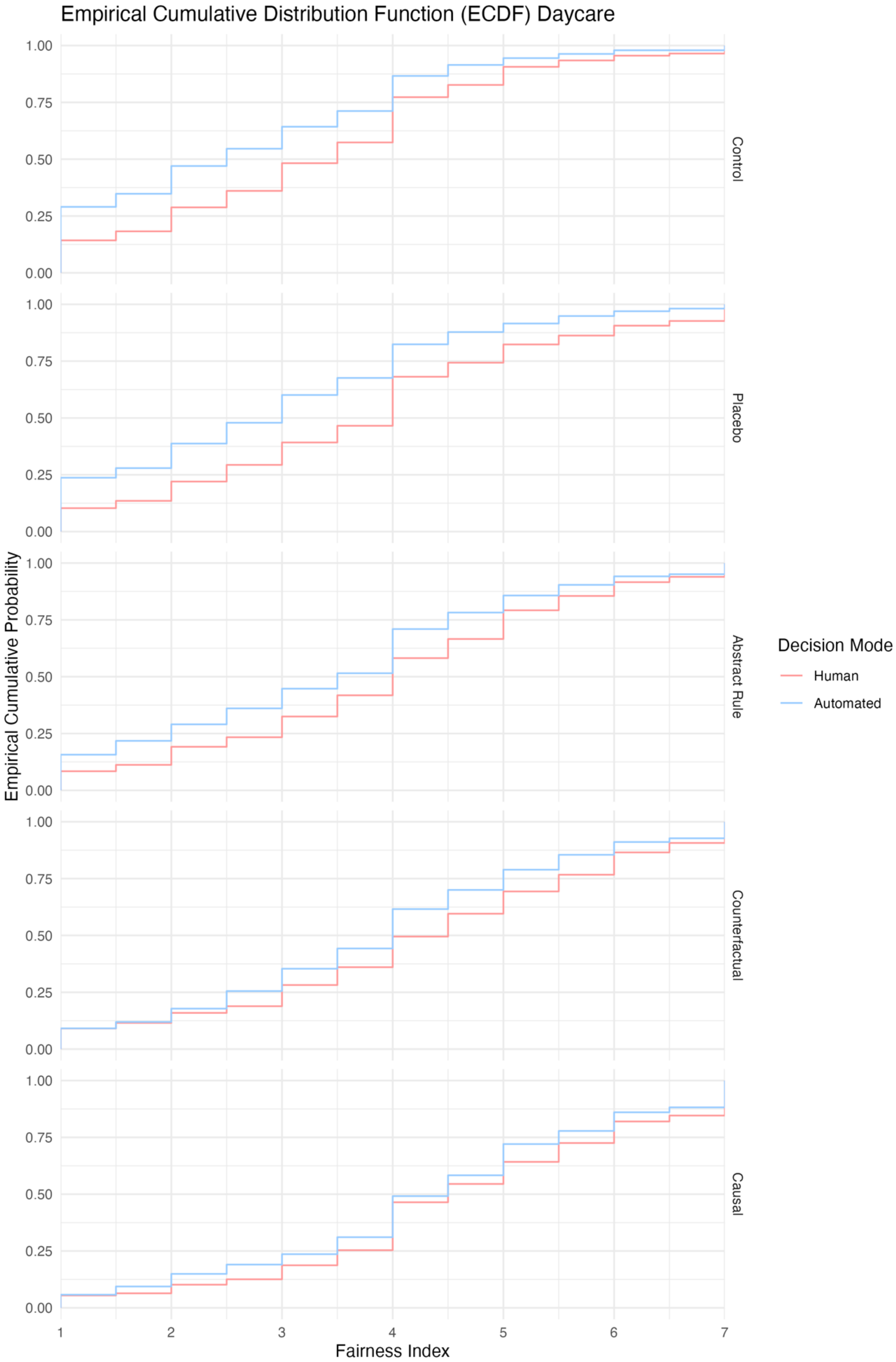
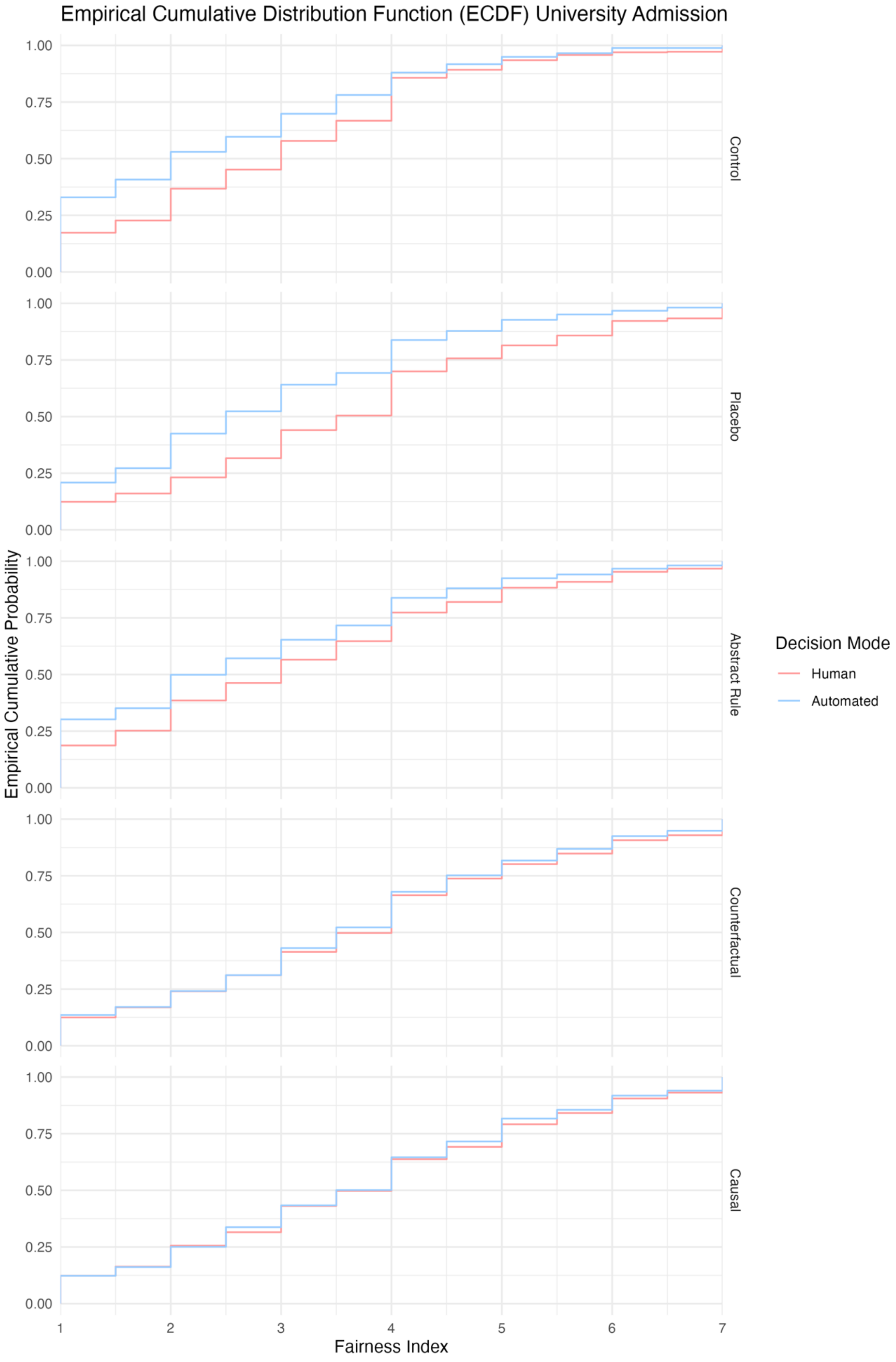


Figure D3: Empirical Cumulative Distribution Functions of Fairness Index by Decision Mode for each Treatment (University Admission)



## Appendix E: Summary Analyses of Different Fairness Measures

Table E1: Replication of Table 3 for outcome and procedural fairness separately

| DV: Fairness of | (1)<br>Outcome | (2)<br>Outcome | (3)<br>Procedure | (4)<br>Procedure |
|---|---|---|---|---|
| PLACEBO | 0.358*** | 0.384*** | 0.341*** | 0.412*** |
| | (0.0642) | (0.0904) | (0.0665) | (0.0938) |
| ABSTRACT RULE | 0.344*** | 0.245*** | 0.477*** | 0.442*** |
| | (0.0643) | (0.0909) | (0.0667) | (0.0942) |
| CAUSAL | 0.944*** | 0.726*** | 1.091*** | 0.918*** |
| | (0.0647) | (0.0911) | (0.0670) | (0.0945) |
| COUNTERFACTUAL | 0.853*** | 0.631*** | 0.996*** | 0.848*** |
| | (0.0647) | (0.0919) | (0.0671) | (0.0953) |
| Automated | -0.369*** | -0.570*** | -0.473*** | -0.584*** |
| | (0.0409) | (0.0905) | (0.0424) | (0.0939) |
| PLACEBO*Automated | | -0.0553 | | -0.144 |
| | | (0.128) | | (0.133) |
| ABSTRACT RULE * Automated | | 0.198 | | 0.0698 |
| | | (0.128) | | (0.133) |
| CAUSAL* Automated | | 0.438*** | | 0.345*** |
| | | (0.129) | | (0.134) |
| COUNTERFACTUAL* Automated | | 0.437*** | | 0.293** |
| | | (0.129) | | (0.134) |
| Daycare | 0.259*** | 0.259*** | 0.123*** | 0.123*** |
| | (0.0315) | (0.0315) | (0.0308) | (0.0308) |
| University Admission | -0.198*** | -0.198*** | -0.350*** | -0.350*** |
| | (0.0315) | (0.0315) | (0.0308) | (0.0308) |
| Demographics | ✓ | ✓ | ✓ | ✓ |
| Constant | 3.045*** | 3.135*** | 3.386*** | 3.433*** |
| | (0.190) | (0.193) | (0.196) | (0.200) |
| Observations | 12,750 | 12,750 | 12,750 | 12,750 |
| Number of groups | 4,250 | 4,250 | 4,250 | 4,250 |

Results from multilevel models run on the full sample of observations. The dependent variable is either outcome fairness or procedural fairness. Observations are grouped at the level of the individual. The reference category for the treatments is the CONTROL treatment, and for the scenarios it is the reallocation-of-refugees scenario. Demographic controls include age, gender, education, and parenthood. Standard errors in parentheses. *** p<0.01

Table E2: Comparing outcome and procedural fairness for each treatment separately (collapsed over scenarios, paired t-tests)

| Treatment/Decision Mode | Human | | | Automated | | |
|---|---|---|---|---|---|---|
| | Outcome | Procedure | t-test,p | Outcome | Procedure | t-test, p |
| Control | 3.22 | 3.22 | = .926 | 2.67 | 2.65 | = .531 |
| Placebo | 3.62 | 3.64 | = .641 | 2.98 | 2.90 | < .05 |
| Abstract Rule | 3.5 | 3.69 | < .001 | 3.10 | 3.15 | = .156 |
| Counterfactual | 3.88 | 4.08 | < .001 | 3.83 | 3.91 | < .05 |
| Causal | 3.94 | 4.13 | < .001 | 3.73 | 3.79 | = .138 |