MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

ARIAN HENNING
PASCAL LANGENBACH

Discussion Paper
2024/11

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

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

by

Arian Henning and Pascal Langenbach*

This version May 30, 2025

Abstract

Automated legal decision-making is often perceived as less fair than its human counterpart. This human-AI 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, sufficiently individualized reasoning models have a stronger positive impact on the perceived fairness of automated decisions than on the perceived fairness of human decisions. This largely mitigates the human-AI fairness gap. The results thus suggest that well-designed reasons can improve the acceptability of automated governance.

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 (McConvey et al., 2023), 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 &

* Arian Henning (corresponding author): Research Fellow, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Email: arian.henning@law-school.de; Pascal Langenbach: Senior Research Fellow, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Data necessary to replicate the results of this article are available upon request from the corresponding author.

Mattiuzzo, 2022; Wu, 2023). A 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 (Chen et al., 2022; Grimmelikhuijsen, 2023; Hermstrüwer & Langenbach, 2023; Wang, 2018). Chen et al. (2022) aptly coined the term 'human-AI fairness gap' for this effect, a phrase that was later referenced by Chief Justice John Roberts in his 2023 year-end report on the federal judiciary (Roberts, 2023, p. 6). The fairness gap between human and automated public decision-making poses two practical problems (see Hermstrüwer & Langenbach, 2023): 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. People's trust and 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). 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).

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-AI fairness gap in public decision-making. The reasoning models differ in their information density, varying from reasoning without substantial information gain to reasoning by explaining the abstract decision criteria and finally more sophisticated reasoning models such as counterfactual or causal explanations. 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 and whether these elements can mitigate the fairness gap between human and automated decision-making (Chen et al., 2022, p. 131). Increasing human oversight of automated public decision-making can enhance perceived fairness (Hermstrüwer & Langenbach, 2023; Kern et al., 2022), as can the implementation of hearing rights (Chen et al., 2022) and providing more comprehensive details about the decision process (Chen et al., 2022; Grgić-Hlača et al., 2018; Grimmelikhuijsen, 2023; Kizilcec, 2016; Lee et al., 2019). While much of the previous research has relied on U.S.-based samples (e.g., Chen et al., 2022; Hermstrüwer & Langenbach, 2023; Grgić-Hlača et al., 2018; Kizilcec, 2016), our study

expands the evidence base by employing a largely representative sample of the German population to explore fairness perceptions of automated decision-making (cf. also Kern et al., 2022).

Using data from an online vignette experiment with 4,250 participants, we first replicate the overall human-AI fairness gap in 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-AI fairness gap. Our results suggest that well-crafted, individualized explanations may not only enhance fairness perceptions in public decision-making but may also serve as a targeted interventions to address the deficit of perceived fairness often associated with automation.

Algorithmic Fairness in Public Decision-Making

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; 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 the decision-making procedures (Lind & Tyler, 1988; Thibaut & Walker, 1975; Tyler, 2006).

In experimental settings, humans often serve as a normative reference for the perception of algorithmic agents. 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). Survey experiments inquiring into the perceived appropriateness, trust, and fairness of decision-making by either human or algorithmic agents usually report a preference for human-centered decision-making (Bigman & Gray, 2018; Castelo et al., 2019; Chen et al., 2022; Kern et al., 2022; Hermstrüwer & Langenbach, 2023; Madhavan &

Wiegmann, 2007). Yet, several studies in the context of legal and public decision-making show that, under specific circumstances, automated decision-making might also be 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 where automated systems are already in use or nearing implementation: the reallocation of refugees across the country, the allocation of child daycare places, and university admissions.

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-AI 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 (Kesari et al., 2024). In this study, we investigate four different reasoning models. 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 86 (1) of the EU AI-Act, for example, gives the persons affected the right in principle to request a ‘clear and 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.

[1] Additionally, see Article 13, Recital 72.

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 are 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. Moreover, 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 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 perceived fairness. Consequently, our study does not implement a counterfactual explanation for an actual automated 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 neither allows 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, nor 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 between the reasoning models and the decision modes, are largely explorative in our study. Finally, although several previous studies reported differences in the fairness evaluations of algorithmic (public) decision-making depending on the decision-context (Grimmelikhuijsen, 2023; Hermstrüwer & Langenbach, 2023; Starke et al., 2022), we

did not have ex-ante hypotheses regarding the perceived fairness in our specific scenarios, nor did we find prior evidence for differential effects of the reasoning-models in the three different contexts of public decision-making. As a result, our reporting of scenario-specific effects is also exploratory in nature.

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

This section explains the data collection, the experimental design, including information on the decision scenarios, the reasoning models, and the decision modes, as well as the measures used in our study. The main hypotheses were preregistered at [https://aspredicted.org/rjzz-fs82.pdf].

Data collection

We collected 4,250 observations. Participants were sourced from *Bilendi*, ensuring representativeness on quotas for age (in the age range of 18–69 years), gender, and education, modeled on the German population. Our participants were roughly evenly distributed across experimental conditions (408–436 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, each consisting of multiple-choice questions related to the content of our vignettes. Participants who failed one of the attention checks were excluded from the study.

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.

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 in which our participants live. In each of the three real-world decision contexts, the potential for automation is already established (Bansak et al., 2018; Caspari et al., 2023; McConvey 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.[3] 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 to our participants in randomized order:

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.

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.

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.
[3] One could argue that participants' beliefs about the technology used for automation could influence their evaluation of the reasoning models, and that information about these models might in turn influence participants' beliefs about the technology used. While we cannot entirely rule this out, we consider it a minor concern for two main reasons. First, most likely detailed knowledge about how (machine learned) models differ in explainability is not widespread among the general public. Second, not informing participants about the specific technology used can be seen as externally valid, as most people interacting with automation in public decision-making will also be unaware of the exact technology behind it. If they do care, they will have to form beliefs based partly on the decisions and explanations they receive. Therefore, if any distortions affect evaluations in our experiment, they are also likely do so in real-world settings.

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 reasoning-models as treatments (CONTROL, RULE-BASED, ABSTRACT CRITERIA, CAUSAL, COUNTERFACTUAL). *Table 1* displays an example for the different treatments in one decision scenario.[4] Roughly, our explanatory models can be divided into three groups:

No explanation (CONTROL)

This treatment serves as the baseline condition. It provides participants with the decision outcome only and does not present any additional information or justification.

Explanations without individual case assessment (RULE-BASED, ABSTRACT CRITERIA)

RULE-BASED

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 lacks any meaningful substantive explanation of the decision, it does indicate that the decision is rule-based, which may underline the claim not to have acted arbitrarily. Since this is a matter of course in the application of administrative law, the reasoning model could, however, also be seen as related to more or less 'empty' justifications using placebic information (cf. Eiband et al., 2019; Langer et al., 1978; Slugoski, 1995).

ABSTRACT CRITERIA

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. While this treatment contains more information than the RULE-BASED treatment, it lacks an individualized assessment of the case.

[4] The wording of our treatments in all decision scenarios can be found in *Table A2* in Appendix A.

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).

COUNTERFACTUAL

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

CAUSAL

In our CAUSAL treatment, participants receive the decision outcome, the 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).

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. Moreover, no signature is provided.[5] By testing the different reasoning models not

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

only for 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).

Table 1: Treatments in the daycare scenario:

<table>
<tr><td>Control</td><td>Rule-Based</td><td>Abstract Criteria</td><td>Counterfactual</td><td>Causal</td></tr>
<tr><td colspan="5">Dear Ms. L,
Unfortunately, we are unable to offer you a place in a bilingual daycare center for your child.</td></tr>
<tr><td></td><td>The allocation of daycare places is based on the applicable regulations for the allocation of places in daycare facilities.</td><td colspan="3">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.</td></tr>
<tr><td colspan="3"></td><td>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.</td><td>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.</td></tr>
</table>

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. One measure was designed to assess the participants' perceived fairness of the decision made (outcome fairness). Another measure was centered on the participants' perceptions of the appropriateness of the applicants' treatment (procedural fairness). Outcome fairness and procedural fairness are highly correlated in our sample ($r = .88$). We use the average ratings of these two fairness measures to create a composite 'Fairness Index'. For ease of reporting, and although not preregistered, we generally refer to this 'Fairness Index' when presenting our results. Results remain largely consistent when the two fairness measures are analyzed separately. Qualitative divergences from the 'Fairness Index' results, where meaningful, are recognized throughout the paper. A summary analysis of

the separate fairness measures is provided in Appendix E. In addition to the fairness measures, we also assessed participants' reports on how understandable the decision in question was.[6] Moreover, in a post-experimental questionnaire, we gathered data on participants' experiences with and knowledge of the different decision scenarios, their evaluations of automation and the administration in general, as well as sociodemographic information.[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.

Human-AI Fairness Gap

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-AI 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-AI 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, RULE-BASED, and ABSTRACT CRITERIA, $p < .001$, COUNTERFACTUAL, $p = .021$, CAUSAL, $p = .059$), which supports our first hypothesis.[9]

---

[6] We solely focus on the fairness measures. However, understandability also correlates highly with the fairness measures ($r > .85$).

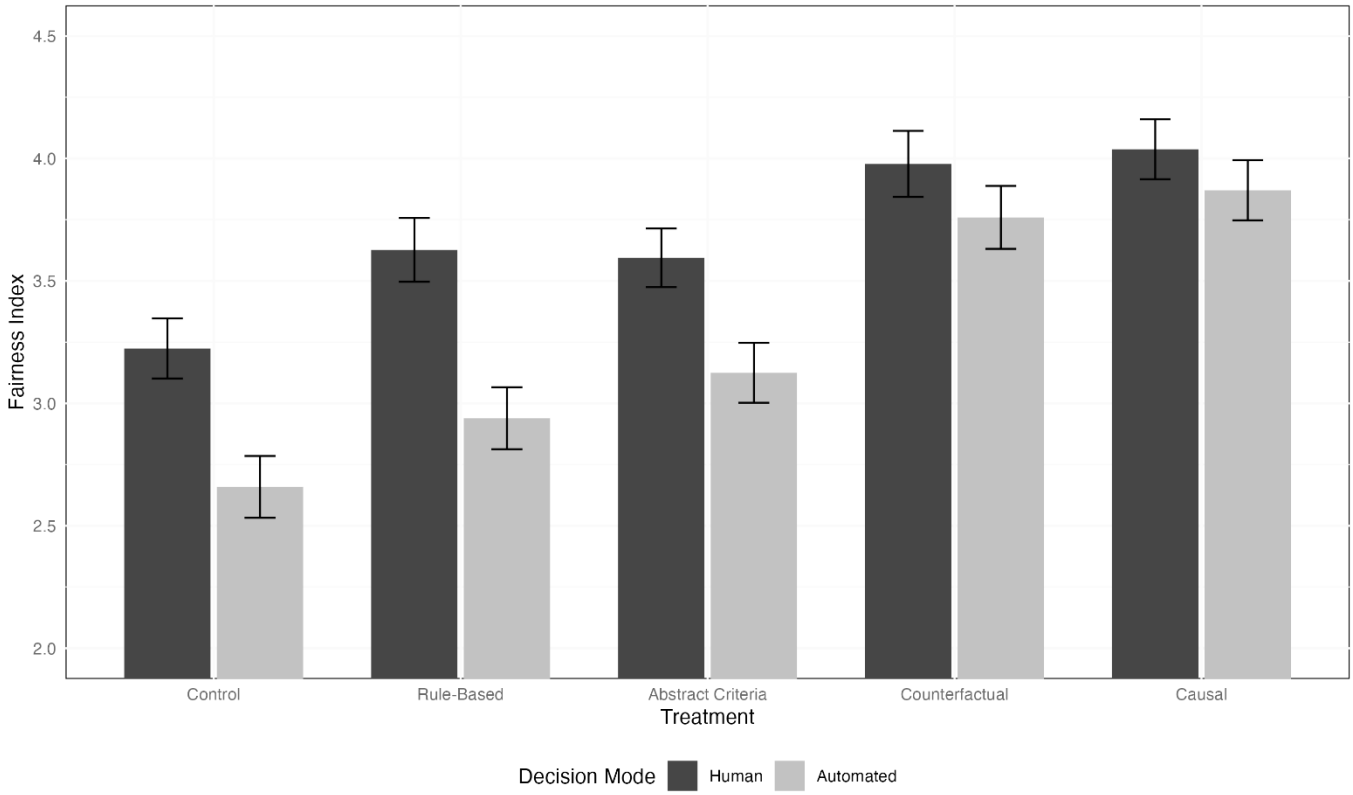
[7] A list of the questions participants were asked can be found in Appendix B.

[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.

[9] Comparing treatments for the two fairness measures separately, the human-AI fairness gap is present for all comparisons when we look at procedural fairness. Outcome fairness, however, is not statistically different in the COUNTERFACTUAL ($p = .141$) and CAUSAL ($p = .223$) treatments between the two decision modes.

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 the 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, $p < .01$).[10] The perceived fairness differences between the CAUSAL/COUNTERFACTUAL treatments and the CONTROL treatment gather 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 the more abstract explanations in the ABSTRACT CRITERIA treatment or the essentially non-

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

informative explanations in the RULE-BASED treatment (comparing CAUSAL | COUNTERFACTUAL vs. ABSTRACT CRITERIA | RULE-BASED, $p < .001$).

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

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: $p = .22$). Differences between the two abstract reasoning models, that is, between the RULE-BASED and the ABSTRACT CRITERIA treatments, are statistically significant in the automated decision mode but not in the human decision mode (Automated: $p = .039$, Human: $p = .717$).[11]

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 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 respondents' 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 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).

[11] However, the treatment difference in automated decision-making is only present for procedural fairness ($p = .009$), but not for outcome fairness ($p = .173$).

Table 2: Fairness differences in reasoning models

	(1)	(2)
DV: Fairness	Human	Automated
Rule-Based	0.391***	0.297***
	(0.0891)	(0.0891)
Abstract Criteria	0.342***	0.480***
	(0.0896)	(0.0890)
Counterfactual	0.737***	1.100***
	(0.0906)	(0.0890)
Causal	0.822***	1.219***
	(0.0898)	(0.0896)
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-AI 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 largely close the fairness gap between human and automated decision-making.

As can be seen from *Figure 1*, automated decisions accompanied by individualized reasons (causal or counterfactual) are descriptively rated as fairer than human decisions with no or only formal reasons. The fairness differences are statistically significant when comparing treatment CAUSAL in the automated decision mode with the treatments without individualized reasons in the human decision mode (CONTROL, RULE-BASED, and ABSTRACT CRITERIA, $p < .01$ for all

pairwise comparisons). Perceived fairness in the COUNTERFACTUAL treatment in the automated decision mode is (marginally) significantly higher than fairness in the CONTROL ($p < .001$) and ABSTRACT CRITERIA treatment ($p = .065$),[12] but not higher than fairness in treatment RULE-BASED ($p = .156$) in the human-decision mode. This indicates that for human decisions which are not explained in a sufficiently individualized way, the human-AI fairness gap can be closed by providing better, that is, more individualized, reasons for automated decisions.

Moreover, providing individualized reasons also affects the human-AI fairness gap when the quality of explanations for both human and automated decisions are equivalent. While fairness in human decision-making dominates fairness in automated decision-making within all reasoning models, the differences between the 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 | 0.22; for all other treatments: > .47).[13] 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*.[14]

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.

---

[12] The difference between the COUNTERFACTUAL and ABSTRACT CRITERIA treatment is significant for outcome fairness ($p = .012$) and insignificant for procedural fairness ($p = .287$).

[13] As reported earlier, we do not find differences between human and automated decision-making for outcome fairness in the individualized treatments. Therefore, for the CAUSAL and COUNTERFACTUAL treatments, there is no evidence that the human-AI fairness gap persists for outcome fairness even in the aggregate data. See below for the scenario specific results.

[14] 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

	(1)	(2)
DV: Fairness		
Rule-Based	0.350***	0.398***
	(0.0632)	(0.0891)
Abstract Criteria	0.411***	0.343***
	(0.0634)	(0.0895)
Counterfactual	0.925***	0.739***
	(0.0637)	(0.0905)
Causal	1.017***	0.822***
	(0.0637)	(0.0898)
Automated	-0.421***	-0.577***
	(0.0403)	(0.0892)
Rule-Based*Automated		-0.0996
		(0.126)
Abstract Criteria*Automated		0.134
		(0.126)
Counterfactual*Automated		0.365***
		(0.127)
Causal*Automated		0.391***
		(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

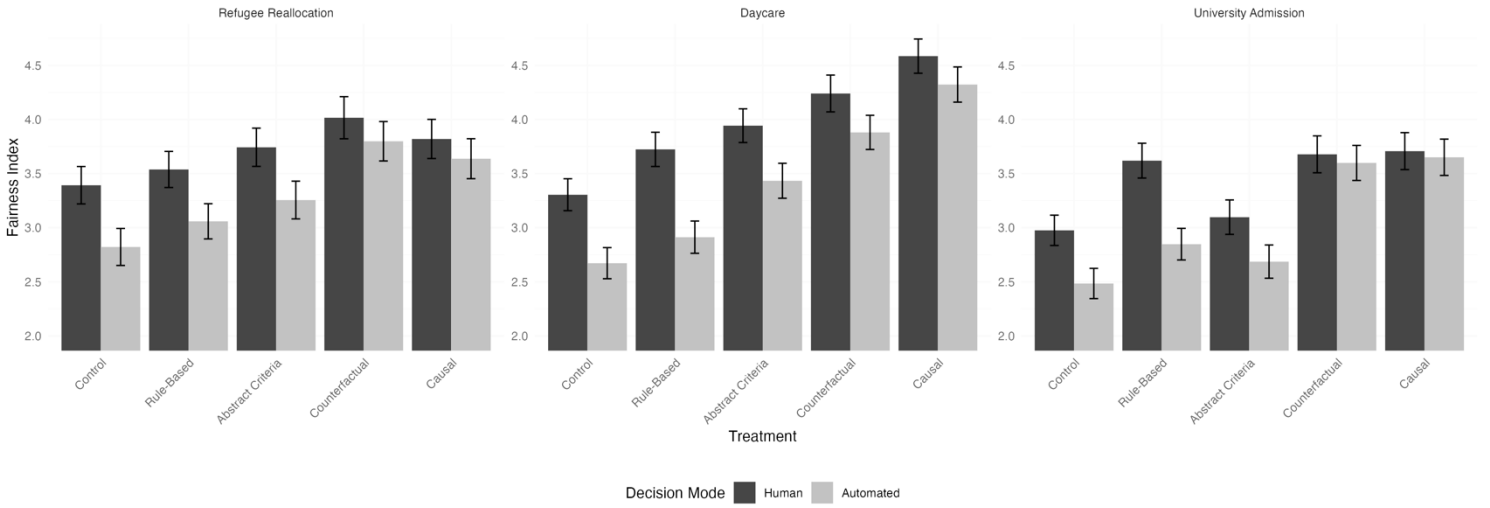
Results from multilevel models run on the full sample of observations. Observations are grouped at the level of the individual. The reference category for the treatments is the CONTROL treatment, and for the scenarios it is the reallocation-of-refugees scenario. Demographic controls include age, gender,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

Context-Specific Effects and Demographics

Previously, we presented 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. As a robustness check for the aggregate results, we provide a breakdown of the differences between the treatments and decision modes for each

of the three scenarios separately. We also include some of the questionnaire data in these analyses. 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 some differences.[15]

Refugee Reallocation – In the refugee-reallocation scenario, while decision-making with giving reasons is perceived overall as fairer than decision-making without giving reasons (CONTROL vs. all else, except RULE-BASED, $p < .01$), the participants do not perceive the RULE-BASED treatment as significantly fairer than the CONTROL treatment ($p = .234$) in the human decision-mode. Yet, in the automated decision-mode it is ($p = .049$). In the human decision mode, the ABSTRACT CRITERIA treatment is not perceived as significantly less fair than the CAUSAL treatment ($p = .550$), 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 ratings of causal and counterfactual explanations are not different though (Human: $p = .147$; Automated: $p = .222$).

Daycare – In the daycare scenario, decision-making accompanied with reasons is generally perceived as fairer than decision-making without giving reasons (CONTROL vs. all else, $p < .03$).

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

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 this scenario, the ABSTRACT CRITERIA treatment is marginally significantly rated as slightly fairer than the RULE-BASED treatment also in the human decision mode ($p = .053$).[16] The differences between the CAUSAL and the COUNTERFACTUAL treatment turn out significant in both decision modes in the daycare scenario (Human: $p = .004$, Automated: $p < .001$).

University Admission – Decision-making with reasons is mostly perceived as fairer than decision-making without reasons (CONTROL vs. all else, except ABSTRACT CRITERIA, $p < .001$). However, in this scenario, the difference between the CONTROL treatment and the ABSTRACT CRITERIA treatment is only marginally significant in the automated decision mode ($p = .056$) and not statistically significant in the human decision mode ($p = .260$).[17] Moreover, under human decision-making, the RULE-BASED explanation leads to a higher fairness score than the ABSTRACT CRITERIA treatment ($p < .001$). In the automated decision mode, there is no statistical difference between the two treatments ($p = .135$).[18] 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 CRITERIA treatment in the human decision mode ($p < .001$). The treatments with individualized reasons are not perceived as fairer than the RULE-BASED treatment in the human decision-mode ($p > .46$). The causal and the counterfactual reasoning models do not receive different fairness ratings either ($p > .66$).

Human AI-Fairness Gap

We find the human-AI fairness gap in the CONTROL treatment (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. In the university-admission scenario, the difference is even descriptively neglectable (CAUSAL: diff = .06, $p = .642$; COUNTERFACTUAL: diff = .08, $p = .507$). In the

[16] This difference is significant for the procedural fairness measure ($p = .024$), but insignificant for the outcome fairness measure ($p = 0.154$).

[17] In both decision modes, for outcome fairness, there are no statistically significant differences between CONTROL and ABSTRACT CRITERIA, whereas differences in procedural fairness turn out to be (marginally) statistically significant (Human: $p = .077$, Automated: $p = .015$). These results are confirmed for the whole sample if we estimate the regression model underlying *Table 4* for procedural and outcome fairness separately.

[18] This time, there is a significant difference for the outcome fairness measure ($p = .014$), but not for the procedural fairness measure ($p = .699$).

refugee-reallocation scenario, the differences remain descriptively more pronounced and approach the marginal significance threshold (CAUSAL: diff = .18, p = .166; COUNTERFACTUAL: diff = .22, p = .109).[19] 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; RULE-BASED: diff = .81; ABSTRACT CRITERIA: 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 find that higher age is correlated with higher fairness ratings in two of the scenarios, 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 leads to higher fairness ratings across all decision contexts, while domain experience leads to lower fairness ratings.

Table 4: Demographics

DV: Fairness	(1) Refugee Reallocation	(2) Daycare	(3) University Admission
Rule-Based	0.210**	0.336***	0.510***
	(0.0878)	(0.0788)	(0.0786)
Abstract Criteria	0.400***	0.693***	0.166**
	(0.0880)	(0.0790)	(0.0788)
Counterfactual	0.800***	1.061***	0.904***
	(0.0885)	(0.0794)	(0.0793)
Causal	0.629***	1.461***	0.957***
	(0.0884)	(0.0794)	(0.0792)
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)

[19] For both treatments, differences in procedural fairness remain indeed (marginally) significant (CAUSAL: p = .086; COUNTERFACTUAL: p = .034).

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	2.748***	2.773***	2.818***
	(0.265)	(0.238)	(0.237)
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. The domain knowledge questions were answered on a scale of 1 to 7, with higher values indicating more knowledge. Domain experience equals 0 if participants or people close to them had no experience in the respective domain, and 1 if they had experience. 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 calls for increased transparency (cf. Olsen et al., 2019). In various regions, most prominently in the European Union, automated decision-making systems are subject to transparency regulations.[20] Transparency is 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). Yet, empirical research indicates that providing more technical information is just one of several paths toward refining

[20] See, e.g., Art. 15 (1) lit. h, Art. 22 GDPR; for high-risk AI applications, see Art. 86 (1), 13, 6 EU AI Act.

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 can reduce trust in automated decisions (Kizilcec, 2016). Therefore, an alternative approach to enhance trust and acceptance might involve explanations understood as communicative processes rather than the mere provision of information (Esposito, 2022b, 2022a).

We have provided experimental evidence on the perceived fairness of human and automated public decision-making under different forms of explanations. The overall finding that human decision-making is perceived as fairer than automated decision-making is in line with the existing literature that often highlights the human-AI 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 the gap. Notably, in some decision contexts, individualized reasons even make the fairness difference practically disappear.

Yet, some studies also report higher fairness ratings for automated decision-making compared to human decision-making in certain scenarios, which contrasts with our findings. Of course, while some of these empirical discrepancies might be just noise, it seems more likely, although in the end speculative, that perceptions of fairness of automated decision-making depend not only on the (policy) context, but also on the particular framing in which automated procedures are implemented (e.g., Hou & Jung, 2021). Even in our data, individualized reasons seem to largely close the human-AI fairness gap. It is therefore plausible that certain implementations of automation might exist that will lead the automated procedure to be even viewed as fairer than human decision-making. For instance, Hermstrüwer and Langenbach (2023) find that if automation is always accompanied by full human decision-making, the resulting hybrid procedure can be rated as fairer than the solely human decision-making. The results in Miller and Keiser (2021) suggest that automation could be preferred as a counterbalance to allegedly discriminatory human decision-making.

Regarding the different reasoning models, we show that any kind of reason-giving enhances the perceived fairness of both human and automated decision-making. This result reinforces the importance of reason-giving in both human and automated 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.[21] Even a largely 'empty' explanation proves to be more effective than providing no explanation whatsoever. This aligns with the 'mindlessness' hypothesis, as introduced by Slugoski (1995), which suggests that people process reason-giving heuristically rather than reflectively when presented with formalized, but non-substantive justifications. Independent of the empirical 'success' of such explanations, of course, normatively the state's obligation to provide justifications should not devolve into a race to the bottom. Merely 'placebic' reasoning models risk replacing genuine transparency with mediated and potentially manipulated explanations, undermining accountability and perpetuating algorithmic opacity (Busuioc et al., 2023).

The differences in fairness perceptions across reasoning models provide insights for the design of automated decision-making systems. Our findings imply that the acceptability of these systems by the public can be enhanced by integrating detailed reasoning. Individualized explanations demonstrate a stronger impact on fairness perceptions in automated decision-making compared to human decision-making. To the best of our knowledge, experimentally showing this interaction effect is genuinely novel within the literature on the fairness of automated public and legal decision-making. However, some previous studies aimed in a similar direction. Particularly, Chen et al. (2022) report positive fairness effects of labeling decision-making as more 'interpretable', and test for a differential effect for human and automated decision-making. Yet, they do not report any significant differences in the impact of interpretability on perceived fairness between the two decision modes. This discrepancy between their and our findings can potentially be attributed to the distinct implementations of interpretability and explainability. Both concepts are theoretically closely connected (Esposito, 2022a), however, the way in which they are operationalized might play a critical role. Chen et al. (2022) conceptualize interpretability as a binary variable, categorizing decisions as either 'interpretable' or 'uninterpretable'. Interpretability moreover emphasizes 'how the outcome is derived, not the provision of a reason for the outcome' (Chen et al., 2022, p. 145). In contrast, our study focuses on the effects of specific reasoning models, each designed to justify a particular decision in the context of public decision-making. While Chen et al. (2022) measure the effect of whether a decision is *potentially* open to explanation, we measure the effect of concrete explanations themselves. It is conceivable that specific explanations of a decision

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

reduce internalized reservations toward automated decision-making more strongly than merely asserting that a decision is in principle open to explanation. Thus, explainability could specifically target the fairness deficit of automated decisions, whereas interpretability could address fairness deficits in decision-making more broadly.

We did not have formal hypotheses regarding potential differences between the individualized reasoning models (causal and counterfactual) in the two decision modes. However, on the one hand, it is plausible that counterfactual explanations perform at least as well as causal explanations in terms of perceived fairness, as they can provide a higher degree of actionability for individual decisions if applied correctly (Poyiadzi et al., 2020). In line with this, Warren et al. (2023) report higher trust judgments of counterfactual than causal explanations. On the other hand, causal explanations could be rated as fairer than counterfactual explanations as people might be more familiar with these types of explanations and inherently associate reason-giving with causality (Pearl & Mackenzie, 2018; Tilly, 2006). Our results are inconclusive in this regard. We did not find significant differences between the two reasoning models in two out of three decision-making contexts for both decision modes. In the daycare scenario causal explanations are rated as the fairest in both decision modes.

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 technically 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. 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 poses 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. Consequently, we designed reasoning models that capture core features of the different explanation styles, but were equally applicable to both human and automated 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. However, although our data does not allow us to test this conjecture, the fairness effects of explanations might, in turn, depend on a certain level of complexity and ambiguity in the decision criteria and facts of the case. In the inverse setting, where decision criteria and facts are fully known ex-ante, explanations may add little informational value. In such cases, individuals could rely on the pre-known criteria and their application to the facts to form fairness judgments, without needing additional explanations. 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 thank Léon Bartosch for implementing the survey in Qualtrics and Brian Cooper for language help. We are further very grateful for helpful comments by two anonymous reviewers and 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 when presenting in seminars at ETH Zurich, Humboldt University of Berlin, and the Max Planck Institute for Research on Collective Goods in Bonn; as well as at the CELS Global Conference 2024 in São Paulo, and the ICON-S Conference 2024 in Madrid.

VI. References

Almeida, Denise, Konstantin Shmarko, and Elizabeth Lomas. 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): 377–87. https://doi.org/10.1007/s43681-021-00077-w.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400829828.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence." *AI & Society* 35 (3): 611–23. https://doi.org/10.1007/s00146-019-00931-w.

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. 2018. "Improving Refugee Integration through Data-Driven Algorithmic Assignment." *Science* 359 (6373): 325–29.

Bawden, David, and Lyn Robinson. 2020. "Information Overload: An Overview." *Oxford Research Encyclopedia of Literature.* Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1360.

Bigman, Yochanan E., and Kurt Gray. 2018. "People Are Averse to Machines Making Moral Decisions." *Cognition* 181 (December): 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Article 377, 1–14. New York: Association for Computing Machinery. https://doi.org/10.1145/3173574.3173951.

Busuioc, Madalina, Deirdre Curtin, and Marco Almada. 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.

Caspari, Gian, Sylvia Greshake, Thilo Klein, and Tobias Riehm. 2023. "Software für eine schnelle, faire und transparente Kitaplatzvergabe." https://kitamatch.com/assets/KitaMatch_Manual.pdf.

Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann. 2019. "Task-Dependent Algorithm Aversion." *Journal of Marketing Research* 56 (5): 809–25. https://doi.org/10.1177/0022243719851788.

Chen, Benjamin Minhao, Alexander Stremitzer, and Kevin Tobia. 2022. "Having Your Day in Robot Court." *Harvard Journal of Law & Technology* 36 (1): 128–169. https://jolt.law.harvard.edu/assets/articlePDFs/v36/Chen-Stremitzer-Tobia-Having-Your-Day-in-Robot-Court.pdf.

Chou, Yu-Liang, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2022. "Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications." *Information Fusion* 81 (May): 59–83. https://doi.org/10.1016/j.inffus.2021.11.003.

Clancey, William J. 1983. "The Epistemology of a Rule-Based Expert System—A Framework for Explanation." *Artificial Intelligence* 20 (3): 215–51. https://doi.org/10.1016/0004-3702(83)90008-5.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (1): 114–26. https://doi.org/10.1037/xge0000033.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In *Proceedings of the 3rd Innovations in Theoretical*

Computer Science Conference, 214–26. New York: Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255.

Eiband, Malin, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. "The Impact of Placebic Explanations on Trust in Intelligent Systems." In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper LBW0243, 1–6. New York: Association for Computing Machinery. https://doi.org/10.1145/3290607.3312787.

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Washington, DC: Administrative Conference of the United States. https://www.acus.gov/report/government-algorithm-artificial-intelligence-federal-administrative-agencies.

Esposito, Elena. 2022a. "Does Explainability Require Transparency?" *Sociologica* 16 (3): 17–27. https://doi.org/10.6092/issn.1971-8853/15804.

Esposito, Elena. 2022b. "Transparency versus Explanation: The Role of Ambiguity in Legal AI." *Journal of Cross-Disciplinary Research in Computational Law* 1 (2). https://journalcrcl.org/crcl/article/view/10.

Gajane, Pratik, and Mykola Pechenizkiy. 2018. "On Formalizing Fairness in Prediction with Machine Learning." *arXiv.* https://doi.org/10.48550/arXiv.1710.03184.

Gerstenberg, Tobias, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. 2021. "A Counterfactual Simulation Model of Causal Judgments for Physical Events." *Psychological Review* 128 (5): 936–75. https://doi.org/10.1037/rev0000281.

Goethals, Sofie, David Martens, and Toon Calders. 2023. "PreCoF: Counterfactual Explanations for Fairness." *Machine Learning,* March. https://doi.org/10.1007/s10994-023-06319-8.

Grgić-Hlača, Nina, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction." In *Proceedings of the 2018 World Wide Web Conference,* 903–12. Geneva: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186138.

Grimmelikhuijsen, Stephan. 2023. "Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making." *Public Administration Review* 83 (2): 241–62. https://doi.org/10.1111/puar.13483.

Grove, William, David Zald, Boyd Lebow, Beth Snitz, and Chad Nelson. 2000. "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12 (1): 19–30. https://doi.org/10.1037/1040-3590.12.1.19.

Gryz, Jarek, and Marcin Rojszczak. 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, Yoan, and Pascal Langenbach. 2023. "Fair Governance with Humans and Machines." *Psychology, Public Policy, and Law* 29 (4), 525–48. https://doi.org/10.1037/law0000381.

Hou, Yoyo Tsung-Yu, and Malte F. Jung. 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. https://doi.org/10.1145/3479864.

Janssen, Marijn, and George Kuk. 2016. "The Challenges and Limits of Big Data Algorithms in Technocratic Governance." *Government Information Quarterly* 33 (3): 371–77. https://doi.org/10.1016/j.giq.2016.08.011.

Jussupow, Ekaterina, Izak Benbasat, and Armin Heinzl. 2020. “Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion.” In *Proceedings of the 28th European Conference on Information Systems (ECIS)*, June 15–17, 2020. AIS Electronic Library (AISeL). https://aisel.aisnet.org/ecis2020_rp/168.

Keil, Frank C. 2006. "Explanation and Understanding." *Annual Review of Psychology* 57 (1): 227–54. https://doi.org/10.1146/annurev.psych.57.102904.190100.

Kesari, Aniket, Daniela Sele, Elliott Ash, and Stefan Bechtold. 2024. *A Legal Framework for eXplainable Artificial Intelligence*. Law, Economics and Data Science Working Paper No. 03. Zurich: ETH Zurich. https://doi.org/10.2139/ssrn.4972085.

Kern, Christoph, Frederic Gerdon, Ruben L. Bach, Florian Keusch, and Frauke Kreuter. 2022. *Patterns* 3 (10): 100591. https://doi.org/10.1016/j.patter.2022.100591.

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. “Avoiding Discrimination through Causal Reasoning.” In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 656–666. Red Hook, NY: Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3294771.3294834.

Kizilcec, René F. 2016. "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 2390–95. New York: ACM.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–93. https://doi.org/10.1093/qje/qjx032.

Komiak, Sherrie Y. X., and Izak Benbasat. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents." *MIS Quarterly* 30 (4): 941–60. https://doi.org/10.2307/25148760.

Langer, Ellen J., Arthur Blank, and Benzion Chanowitz. 1978. "The Mindlessness of Ostensibly Thoughtful Action: The Role of ‘Placebic’ Information in Interpersonal Interaction." *Journal of Personality and Social Psychology* 36 (6): 635–42. https://doi.org/10.1037/0022-3514.36.6.635.

Lee, Min Kyung, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 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. Allan, and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice.* New York: Springer Science & Business Media.

Longoni, Chiara, Andrea Bonezzi, and Carey Morewedge. 2019. “Resistance to Medical Artificial Intelligence.” *Journal of Consumer Research* 46 (4): 629–650. https://doi.org/10.1093/jcr/ucz013.

Madhavan, P., and D. A. Wiegmann. 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, Frank, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. "Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 122–30. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372867.

McCloy, Rachel, and Ruth M. J. Byrne. 2002. "Semifactual 'Even If' Thinking." *Thinking & Reasoning* 8 (1): 41–67. https://doi.org/10.1080/13546780143000125.

McConvey, Kelly, Shion Guha, and Anastasia Kuzminykh. 2023. "A Human-Centered Review of Algorithms in Decision-Making in Higher Education." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* 1–15. New York: Association for Computing Machinery. https://doi.org/10.1145/3544548.3580658.

Mehdiyev, Nijat, Constantin Houy, Oliver Gutermuth, Lea Mayer, and Peter Fettke. 2021. "Explainable Artificial Intelligence (XAI) Supporting Public Administration Processes – On the Potential of XAI in Tax Audit Processes." In *Proceedings of the International Conference on Business Information Systems (BIS 2021),* 413–28. https://doi.org/10.1007/978-3-030-86790-4_28.

Mendes, Laura Schertel, and Marcela Mattiuzzo. 2022. "Algorithms and Discrimination: The Case of Credit Scoring in Brazil." In *Personality and Data Protection Rights on the Internet: Brazilian and German Approaches,* edited by Marion Albers and Ingo Wolfgang Sarlet, 407–43. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-90331-2_17.

Miller, Susan M., and Lael R. Keiser. 2021. "Representative Bureaucracy and Attitudes Toward Automated Decision Making." *Journal of Public Administration Research and Theory* 31 (1): 150–65. https://doi.org/10.1093/jopart/muaa019.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019. "Explaining Explanations in AI." In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 279–88. New York: ACM. https://doi.org/10.1145/3287560.3287574.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. "Definitions, Methods, and Applications in Interpretable Machine Learning." *Proceedings of the National Academy of Sciences of the United States of America* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

Nagtegaal, Rosanna. 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, Ana. 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): 869–91. https://doi.org/10.1108/JD-04-2019-0066.

Neches, R., W. R. Swartout, and J. D. Moore. 1985. "Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development." *IEEE Transactions on Software Engineering* SE-11 (11): 1337–51. https://doi.org/10.1109/TSE.1985.231882.

Nissen, Mark E., and Kishore Sengupta. 2006. "Incorporating Software Agents into Supply Chains: Experimental Investigation with a Procurement Task." *MIS Quarterly* 30 (1): 145–66. https://doi.org/10.2307/25148721.

Olsen, Henrik Palmer, Jacob Livingston Slosser, Thomas Troels Hildebrandt, and Cornelius Wiesener. 2019. "What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration." *SSRN Scholarly Paper.* Rochester, NY. https://doi.org/10.2139/ssrn.3402974.

Önkal, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 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.

Pearl, Judea. 2009. *Causality.* 2nd ed. Cambridge: Cambridge University Press.

Pearl, Judea. 2013. "Structural Counterfactuals: A Brief Introduction." *Cognitive Science* 37 (6): 977–85. https://doi.org/10.1111/cogs.12065.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books.

Pieters, Wolter. 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, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. "FACE: Feasible and Actionable Counterfactual Explanations." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 344–50. New York: ACM. https://doi.org/10.1145/3375627.3375850.

Roberts, John. 2023. "2023 Year-End Report on the Federal Judiciary." *Supreme Court of the United States.* https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf.

Shin, Donghee, and Yong Jin Park. 2019. "Role of Fairness, Accountability, and Transparency in Algorithmic Affordance." *Computers in Human Behavior* 98 (September): 277–84. https://doi.org/10.1016/j.chb.2019.04.019.

Simmons, Ric. 2018. "Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System." *UC Davis Law Review* 52 (2): 1067–1118. https://lawreview.law.ucdavis.edu/archives/52/2/big-data-machine-judges-and-legitimacy-criminal-justice-system.

Slugoski, B. R. 1995. "Mindless Processing of Requests? Don't Ask Twice." *The British Journal of Social Psychology* 34 (Pt 3): 335–50. https://doi.org/10.1111/j.2044-8309.1995.tb01068.x.

Sokol, Kacper, and Peter Flach. 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,* CEUR Workshop Proceedings 2301.

Starke, Christopher, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. "Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature." *Big Data & Society* 9 (October). https://doi.org/10.1177/20539517221115189.

Swartout, William 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, John W., and Laurens Walker. 1975. *Procedural Justice: A Psychological Analysis.* Hillsdale, NJ: L. Erlbaum Associates.

Tilly, Charles. 2006. *Why?* Princeton, NJ: Princeton University Press.

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

Tyler, Tom R. 2006. *Why People Obey the Law.* 2nd ed. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctv1j66769.

Tyler, Tom R., and Yuen J. Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts.* New York: Russell Sage Foundation.

Tyler, Tom R., and Jonathan Jackson. 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, Sandra. 2022. "The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law." *SSRN Scholarly Paper.* Rochester, NY. https://doi.org/10.2139/ssrn.4099100.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *arXiv.* https://doi.org/10.48550/arXiv.1711.00399.

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

Warren, Greta, Ruth M. J. Byrne, and Mark T. Keane. 2023. "Categorical and Continuous Features in Counterfactual Explanations of AI Systems." In *Proceedings of the 28th International Conference on Intelligent User Interfaces,* 171–87. Sydney, NSW: ACM. https://doi.org/10.1145/3581641.3584090.

Wu, Yi. 2023. "Data Governance and Human Rights: An Algorithm Discrimination Literature Review and Bibliometric Analysis." *Journal of Humanities, Arts and Social Science* 7 (1): 128–54. https://doi.org/10.26855/jhass.2023.01.018.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. "Making Sense of Recommendations." *Journal of Behavioral Decision Making* 32 (4): 403–14. https://doi.org/10.1002/bdm.2118.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32 (4): 661–83. https://doi.org/10.1007/s13347-018-0330-6.

VII. 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. On his arrival in Germany, M was assigned to the Cologne-Bayenthal initial reception center. 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, Your application for relocation to the initial reception center in Hamburg-Rahlstedt has been rejected.	Dear Ms. L, Unfortunately, we are unable to offer you a place in a bilingual daycare center for your child.	Dear Ms. A, Thank you for your interest in studying at the University of Düsseldorf. Unfortunately, you cannot be admitted to the Master's program in Psychology.
Rule-Based (Control + 'Empty' Rule)	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 Criteria (Control + Abstract Criteria)	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 (Control + Abstract Criteria + 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 (Control + Abstract Criteria + 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, Meyer	This letter was generated by machine and is therefore valid without a signature.

VIII. 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 willing 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)

Have you personally, or does someone close to you, had 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 admission degree programs?
Yes/No

IX. Appendix C: Post-Regression Wald Tests for Treatment Differences

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

	Human	Automated
Rule-Based vs. Abstract Criteria	p = .581	p = .04
Rule-Based vs. Counterfactual	p < .001	p < .001
Rule-Based vs. Causal	p < .001	p < .001
Abstract Criteria vs. Counterfactual	p <. 001	p < .001
Abstract Criteria vs. Causal	p <. 001	p < .001
Causal vs. Counterfactual	p = .347	p = .186

X. Appendix D: Empirical Cumulative Distribution Functions

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

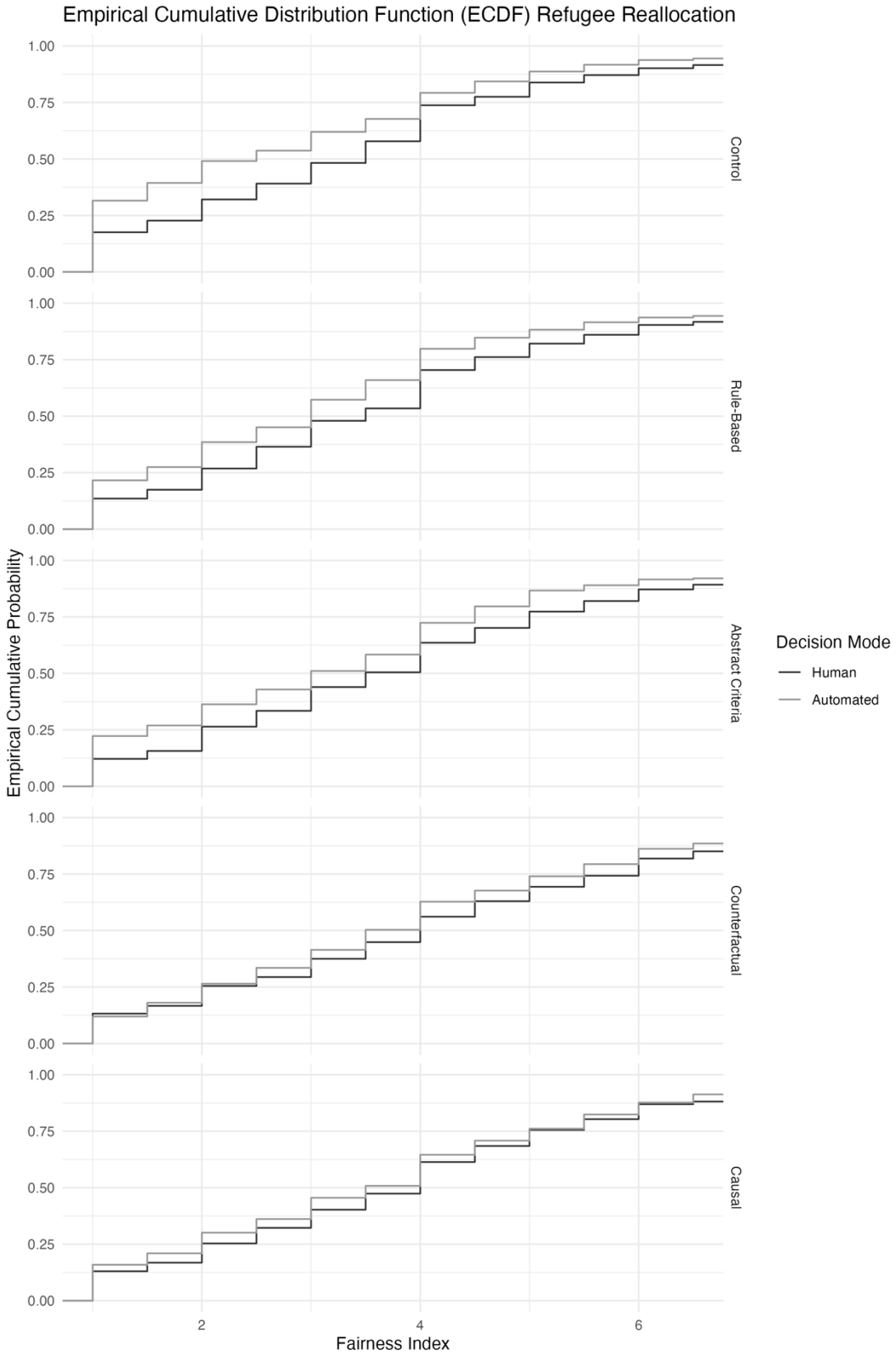


Figure D2: Empirical Cumulative Distribution Functions of Fairness Index by Decision Mode for each Treatment (Daycare)

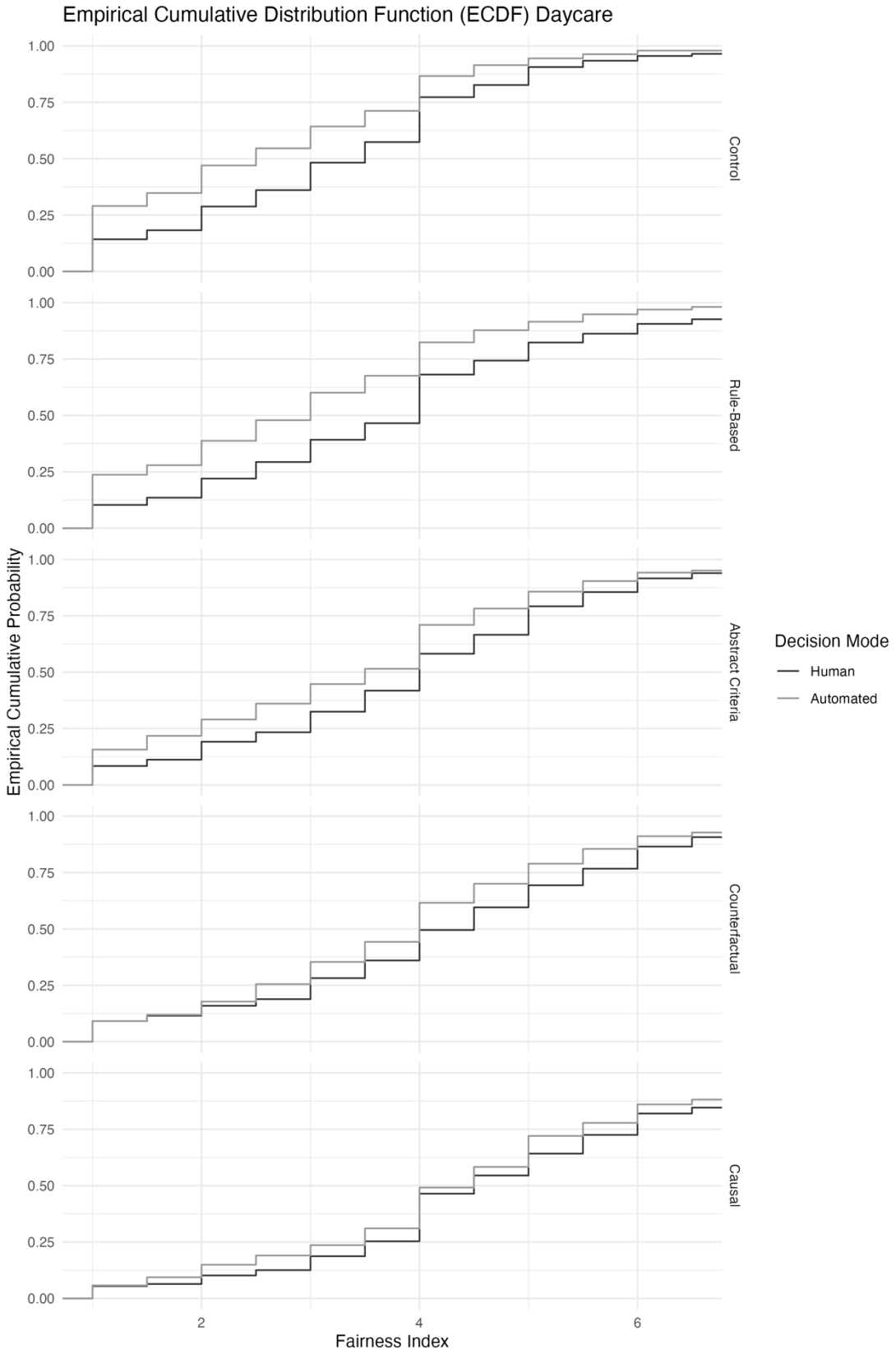
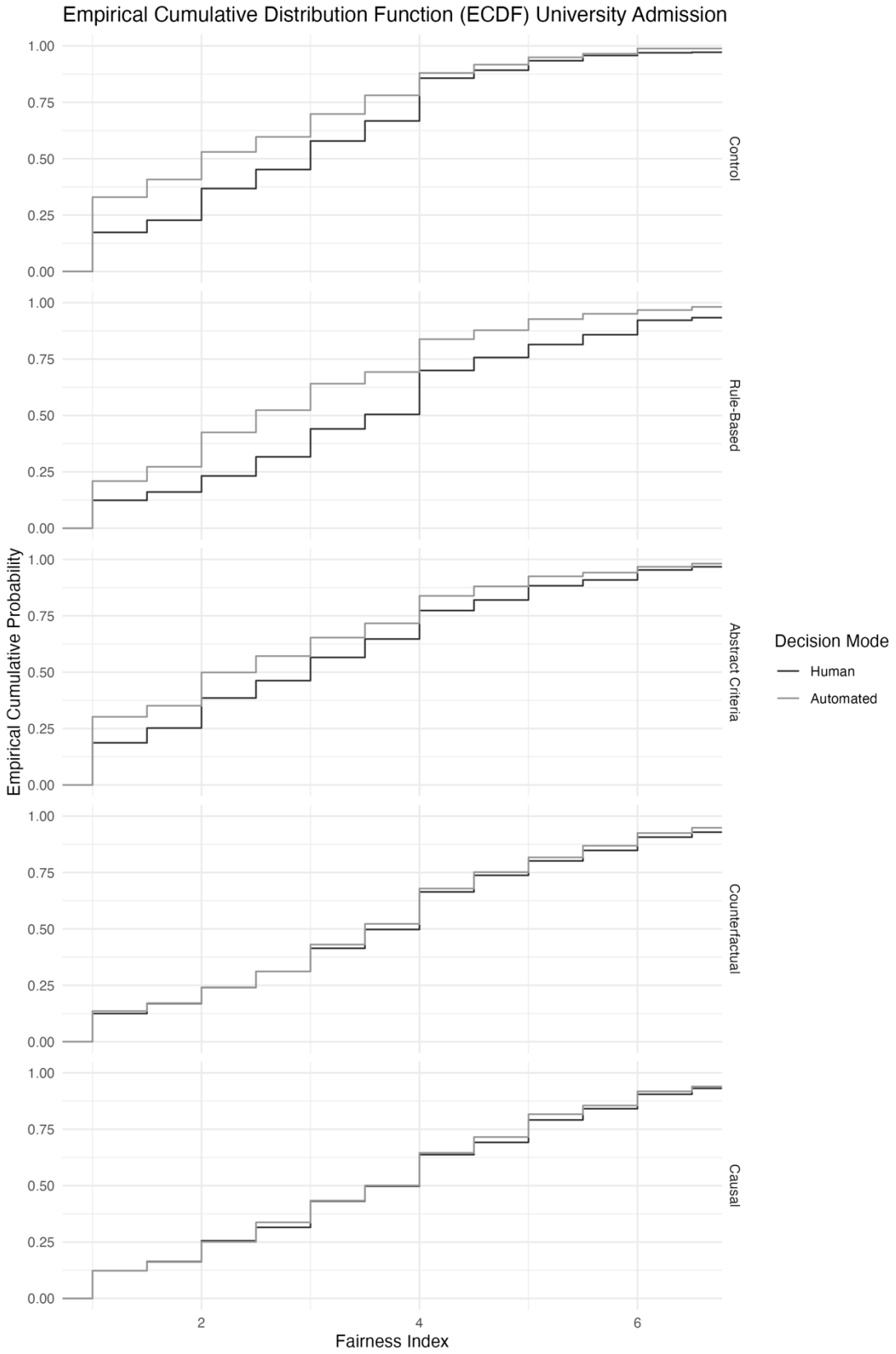


Figure D3: Empirical Cumulative Distribution Functions of Fairness Index by Decision Mode for each Treatment (University Admission)



XI. Appendix E: Summary Analyses of Different Fairness Measures

Table E1: Conceptual replication of *Table 3* for outcome and procedural fairness separately

	(1)	(2)	(3)	(4)
DV: Fairness of	Outcome	Outcome	Procedure	Procedure
Rule-Based	0.358***	0.384***	0.341***	0.412***
	(0.0642)	(0.0904)	(0.0665)	(0.0938)
Abstract Criteria	0.344***	0.245***	0.477***	0.442***
	(0.0643)	(0.0909)	(0.0667)	(0.0942)
Counterfactual	0.853***	0.631***	0.996***	0.848***
	(0.0647)	(0.0919)	(0.0671)	(0.0953)
Causal	0.944***	0.726***	1.091***	0.918***
	(0.0647)	(0.0911)	(0.0670)	(0.0945)
Automated	-0.369***	-0.570***	-0.473***	-0.584***
	(0.0409)	(0.0905)	(0.0424)	(0.0939)
Rule-Based*Automated		-0.0553		-0.144
		(0.128)		(0.133)
Abstract Criteria*Automated		0.198		0.0698
		(0.128)		(0.133)
Counterfactual*Automated		0.437***		0.293**
		(0.129)		(0.134)
Causal*Automated		0.438***		0.345***
		(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	3.23	3.22	= .926	2.67	2.65	= .531
Rule-Based	3.62	3.64	= .641	2.98	2.90	< .05
Abstract Criteria	3.5	3.69	< .001	3.10	3.15	= .156
Counterfactual	3.88	4.08	< .001	3.73	3.79	= .138
Causal	3.94	4.13	< .001	3.83	3.91	< .05