Journel of AI and Machine Learning
Volume 1, Issue 1

SCHOLARLY SUMMIT
THE ACADEMIC PUBLISHERS

Explainable AI in Public Policy: Quantifying Trust and Distrust in Algorithmic Decision-Making across Marginalized Communities

[1]*Saad Tehreem,[2]Hammad Razi

[1], [2] Department of Marketing, College of Management Sciences, PAF-KIET University, Pakistan

Abstract

Explainable AI (XAI) attempts to provide explanations and, thus, increase trust while the effect is influenced by demographic factors, cultural match, and perceived fairness. This study explores the role of trust and distrust regarding sociodemographic data and perceived fairness in the SOC system decision-making. It examines whether cultural match influences perceptions of fairness and whether the nature of the explanation (technical, plain language, or human) influences trust. This study employed a cross-sectional survey of 240 participants and used experimental vignettes where participants received decisions from an AI with/without explanations in one of three types. Relationship perception and algorithmic distrust and trust were analyzed using regression, MANOVA, and mediation. The study shows that human decisions are trusted most and are followed by plain language AI-generated reasons; the least trusted are technical reasons. It was also shown that perceived fairness regulates trust and that low-income users are more sensitive to fairness perception. Culture has been proven to have a strong association with fairness perception, emphasizing the need to adopt the context-based approach for AI governance. However, passive exposure to AI does not imply trust and, therefore, requires that transparency be perceived appropriately by the public. This study aims to expand knowledge in AI governance by attempting to apply both procedural justice and algorithmic accountability frameworks. It points to a lack of generalized public trust in AI and stresses the need for culturally sensitive and inclusive AI designs. Recommendations indicate that explainability is more important than just the technical process to drive policy changes.

Keywords: Explainable AI (XAI), Public Policy, Cultural Alignment, Algorithmic Trust, Experimental Vignettes.

1. Introduction

Artificial intelligence (AI) has become one of the most significant trends in governing public policies in fields such as determined allocation of welfare, police foreseen quantity, and healthcare availability. Governments are increasingly turning to AI to work more effectively and make decisions to improve the distribution of resources and effectiveness of delivered services (Robles & Mallinson, 2023; Kaushik, 2020; Wang et al., 2024). However, the integration of AI into the planning of public policy has sparked important ethical and social issues, especially in terms of the level of equity. Algorithms, when implemented, tend to create conflict between the optimization of tasks and the need for fairness to minority groups (Toll et al., 2020; Papadakis et al., 2024).

Concerns about the effect of AI on justice and equity in society have also increased. AI decision-making processes can be prejudiced, particularly when these algorithms are founded on the current bias and unequal frameworks (Araujo et al., 2020; Misra et al., 2020). For example, Mahsood et al. (2020) explain that it may have negative effects on mitigated groups of people, so there are demands for accountability examining the usage of AI (Esmaeilzadeh, 2020). The problem is the irrational imbalance that has to be addressed to make AI systems both effective and acceptable to the communities in which they operate.

Another vital notion in this context is XAI, or explainable AI that aims at providing rationales behind AI-based decisions to end-users (Amarasinghe et al., 2023). However, the largely technical definitions of explainability do not capture the sociocultural factors that define the

Saad Tehreem
[1]Department of Marketing, College of Management Sciences, PAF-KIET University, Pakistan
[1]*Saadtehreem@researchergroup.co

Received: 02-Jan-2024 Revised: 03-March-2025 Accepted: 10-March-2025

social relations of minority communities and determine how they trust these AI systems (Arrieta et al., 2020; Veer et al., 2021). This lack of awareness remains problematic about trust as this is a requirement for the application of AI in public policy domains (Coyle & Weller, 2020; Zhang & Dafoe, 2020).

Furthermore, the existing literature lacks information on how such vulnerable groups interpret and trust explainable AI tools. Despite its perceptual clarity, the abstract definitions in technical terms of identifying explainability ignore the underlying sociocultural context of fairness that is considered important by the general public (Ploug et al., 2021; Christou et al., 2023). These are issues that should not be witnessed, especially in marginalized groups of people who may be victims of an unjust society. Therefore, an imperative arises to investigate how sociodemographic factors and the perceived fairness of XAI affect trust and distrust in algorithmic decisions in these groups.

Expanding on the above concerns, this study aims to analyze the correlation between the demographic variables, perceived fairness, and trust in Explainable AI in marginalized group populations and answer the following research question:

RQ01: How do sociodemographic factors and perceived fairness of explainable AI systems predict trust/distrust in algorithmic decision-making among marginalized communities?

The study seeks to fill the gap in the literature by offering vital statistical insight into the factors that shape trust and distrust in algorithmic decision-making. In addition, the study identified how gender, income level, and the use of technology influence fairness and trustworthiness assertions of an AI system focusing on marginalized communities (Araujo et al., 2020; Khan & Vice, 2022).

The importance of this research is drawn from the fact that it can contribute towards rethinking and advancing fair policies on AI. Thus, the study is devoted to marginalized groups to add voices to the debates on ethical approaches to AI and politics that involve incorporating minority experiences into algorithmic systems development (Wei et al., 2024; Schiff, 2023). In addition, the findings are useful to policymakers, technocrats, and researchers as they will offer guidelines on how to enhance trust in AI technologies and ways of ensuring AI technologies work for the benefit of society and especially for society's less privileged individuals (Mucci et al., 2024; Bell et al., 2023).

In general, the relationship between technology and trust and equity, as AI is gradually becoming the focus of public policies, should be closely discussed. This will help shed more light on these dynamics and foster the informed use of AI in decision-making among the public.

2. Literature Review

2.1. Trust In Technology: Models Of Human-AI Trust

The idea of trust as a concept in the field of human interactions with intelligent technologies has gained popularity in recent years. Trust is a complex concept that is reflected in the propensity to trust, perceived reliability, and impact factors of personalities that characterize trust. It is important to understand these models because the collaboration between humans and AI is rapidly increasing across various corporations as more systems are being deployed in organizations' decision-making processes.

The extrapolative method identified one of the models of trust as the propensity to trust, which may be defined as an individual's overall level of trust in others, including technology. General trust is actually found to affect AI trust; thus, people with higher trust levels are believed to trust AI systems Montag et al. (2023). For example, Montag et al. (2023) reported that trust in AI is distinct from trust in humans and claimed that trust in such systems involves personality factors in unique ways (Montag et al., 2023). This raises the question of what happens when these personality traits are combined with the perceived reliability of the AI systems.

Perceived reliability is another important aspect of an AI; this is the confidence one has that the integrated AI system will function as expected. Research suggests that users rely on AI-based recommendations more than what is necessary; this is due to overconfidence in AI (Wang et al., 2022; Wang et al. (2022) note that overdependence on AI systems can have adverse effects when users do not adjust their trust level depending on the performance of the AI system (Wang et al., 2022). This points to the need to establish ways and means through which users' reliability assessment of AI is made easier.

In addition, the development of trust in AI has the following dynamics: the context in which it is applied and the user's experience with the application. For instance, Guo et al. (2023) have developed a Trust Inference and Propagation (TIP) that shows how trust develops in MITMRT to highlight the fact that it is not fixed but an interactive process that depends on the interaction with AI (Guo et al., 2023). This model is especially important in stressing the fact that trust must be continually assessed during users' interactions with AI systems while, at the same time, putting forward the idea of reciprocating trust by means of positive experiences and effective communication.

However, despite the above advancement, significant gaps still exist. For example, several studies address the features of the AI system, but there is a scarcity of empirical literature that tackles the cultural facets of trust in AI systems. Trust is not only an assessment of rational judgment but also includes elements of emotion and social influence (Rojas & Li, 2024; Rojas and Li (2024) also highlight other aspects of social considerations for human-AI teams that were highlighted in this article, noting that trust can also be social; that is, it can become infected by interactions within the teams (Rojas & Li, 2024). This not only suggests the need for a new approach to the role of social context and interpersonal relations in trusting AI.

Overall, the models of human-AI trust are full and rich, constituted with individual differences, perceived trustworthiness, and context factors. As AI solutions spread in various domains of users' lives, it is important to examine trust theories to improve HAI cooperation. Further research should be conducted to fill the existing gaps and analyze the relationships between factors belonging to the person, the social environment, and the perceived reliability of AI systems. This will help to create new, more trustworthy, and effective AI technologies.

2.2. Marginalization and Algorithmic Bias

Discrimination and accumulation of injustice in policing, especially in predictive policing systems, have become some of the powerful drivers of injustice towards blacks and other people of color. Predictive policing involves using a model to estimate the likelihood of a crime to occur. This depends on previous data, which still is not free from bias in the police force. For example, the Chicago crime prediction algorithm has been accused of being racially biased by favoring some races over others; hence, it fuels racism and increases giving power to racism (Ziosi & Pruss, 2024; Hung & Yen, 2023). This is because the data has considerable racial bias, which becomes apparent in the evaluation of the risk level of certain neighborhoods, in turn increasing police patrols and, by extension, scrutiny (Susser, 2021; González, 2024).

Research evidence has clearly shown that algorithmic bias has the potential to cause disparities in the way police handle different races. For instance, data proves that Black car owners are pulled over by police more often than White drivers, with the difference depending on the area (Ekstrom et al., 2021; Payne & Rucker, 2022). This is also experienced due to implicit bias from the policemen, whereby the authorities unconsciously have a perception that blacks are criminals, therefore policing them in such a manner (Lai & Lisnek, 2023; Kochel & Nouri, 2024). In addition, the lack of documented evidence on police relations with suspects or black people limits the ability to determine the extent of racial prejudice in the police force since many individuals are not reported if they do not end up under investigation (Knox et al., 2020; Laniyonu & Donahue, 2023).

Such biases are significant because they perpetuate the notion out there that police are partial and unfair, especially to minority groups. Research also establishes that police brutality or racial profiling weakens the level of Confidence in the police among citizens, hence hindering engagement between the two (Andersen et al., 2023; Jimenez et al., 2022). Similarly, the psychological effects of such policing strategies contribute to the increase in the status quo of minority emotions, such as fear and exclusion (Barajas-Gonzalez et al., 2021; Rios et al., 2020).

In general, the intersection of marginalization and algorithmic bias in predictive policing underscores the urgent need for reform in law enforcement practices. Eradicating these biases also entails more than tweaking the predictive models and tweaking the data that feeds the policing strategies. Policymakers require support from society in their endeavors to revive the broken trust and improve the essential aspects of police work across various societies.

2.3. Explainability Metrics - Technical Vs. User-Centric

SHAP values are an example of technical explainability metrics used in the context of feature importance measures in terms of attributions for model predictions. For example, Jeong et al. employed SHAP values to identify the effect of several variables on ion transport across membranes, and some of these features had more profound impacts on the results (Jeong et al., 2023). Similarly, Abbasi et al. have also used SHAP values for the prediction of maize biomass yield and found that growing degree days and cumulative rainfall are the most vital features for model accuracy (Abbasi et al., 2025). These technical features are based on game theory and provide a solid methodological foundation to investigate intricacies present in data (Yan et al., 2020). However, SHAP values are highly informative if their number makes users non-technical and they cannot interpret them well enough.

On the other hand, user-centric interpretability focuses on how the outcomes of the model can be explained to a wide range of users. This approach tries to convey the results in a format that is understandable by laymen most of the time using things like visuals and basic overviews. For instance, Weng et al. stressed the role of governance quality in energy consumption forecasts, noting the need to bring such conclusions into practice for the guidance of policy-makers (Weng et al., 2024). While estimators, such as SHAP, add value in terms of depth, their application poses a risk to users who are only interested in clear and easy-to-comprehend information to aid their decision-making.

However, there is a major research gap in the literature, which lies in the attempt to juxtapose these two approaches. The detailed numeric measures have their advantage and, simultaneously, their limitations from the practical standpoint of governance. On the other hand, personally centered measurements may be more flexible than precise measurements. This could improve the usability of explanations in governance models, where the focus will be on both effectiveness and efficiency for potential end-users. In essence, SHAP values do help technical people understand the performance of a model at the feature level, but decision-makers lack opportunities to use these values in decision-making due to the lack of interpretability.

2.4. Theoretical Framework

The combination of Procedural Justice Theory (PJT) and Algorithmic Accountability ensures that trust of explainable artificial intelligence (XAI) systems within public policy would be effectively interpreted and implemented. This synthesis pertains to the deficiencies of an overly technological framework of fairness addressing issues of sociocultural aspects of trust, as well as power relations within and with regard to algorithmic arrangements.

Procedural Justice Theory (PJT), popularized in legal and organizational psychology, states the organization's credibility is determined by procedural rather than outcome fairness. PJT principles are voice, transparency, impartiality, and respect, which allow for building trust with the communities most affected but who are often not represented. For example, welfare algorithms that do not explain why benefits are being denied can increase distrust, though the results may be seemingly positive (Purves & Davis, 2022). This underlines the need for incorporating practices that enable the community to participate due to exclusion, which creates a feeling of injustice even if the algorithm is right (Baykurt, 2022).

In contrast, Algorithmic Justice seeks the expanse of unfairness contained in AI, thus calling for explainability, auditability, redressal, and responsibility. This framework disapproves of the way that fairness is reduced to measures of statistical parity because people of color do not trust machines, not only because the algorithms are prejudiced but also because of the lack of adequate technical communication (Birhane, 2022). For instance, the use of complex terminologies when explaining certain stages of the algorithm may be incomprehensible for the user due to a lack of background knowledge, which influences perceived fairness (Gursoy & Kakadiaris, 2022).

Additionally, there is a great overlap of the key themes of PJT and Algorithmic Accountability, namely flexibility, social consequences, and complexity. Whereas the PJT focuses on giving mundane and culturally acceptable reasons that are relatable to the users, Algorithmic Accountability poses technical bases such as model understanding (Delgado et al., 2023). Nonetheless, a model that is understandable to engineers may not be acceptable as fair to marginalized users, as revealed by Young et al. (2024). This is because, while PJT encourages the incorporation of minorities in the design of AI systems, Algorithmic Accountability often only allows their input in the form of auditing that tends to be mere 'window dressing.'

From applying this integrated framework to the context of XAI in public policy, it is therefore clear that trust is anchored on procedural justice and structural legitimacy. They must, therefore, be understandable and culturally sensitive to the recipient's dignity. For instance, the eligibility algorithms for healthcare should produce explanations in the users' main language and not use terms that may be unfamiliar (Schmid et al., 2020). In addition, a system must be auditable and co-developed with the marginalized groups as the latter should inform the fairness goals and measurements (Tran & Nguyen, 2020).

However, it has some weaknesses, which are highlighted below. It is also important to note that both PJT and Algorithmic Accountability stem from Euro-American roots and may not be as effective when it comes to collectivist or postcolonial societies (Clarke et al., 2023). Moreover, the overreliance on the XAI can lead to transforming trust as a technical question, which does not take into consideration the formation of distrust among marginalized population groups (Njalsson, 2023).

Therefore, the speculations on the integration of PJT and Algorithmic Accountability present a complex view of trust in AI regarding public policy. It deconstructs fairness in terms of accuracy and puts emphasis on 'distributive fairness,' which makes the voice of the oppressed heard throughout the AI decision-making process and remediation. This is a significant shift in XAI's conceptualization as a sociotechnical system where technical intelligibility should work in tandem with procedural justice for a better overarching governance system.

3. Method and Instruments

This study uses a cross-sectional survey research strategy complemented with integrated experimental scenarios to examine trust in explainable AI (XAI) among marginalized populations. The vignette-based approach enables a measured way of introducing the policy students to AI by providing them with artificial experiences of interactions with the processes governed by algorithms. Using retrospective cross-sectional survey self-assertion combined with quasi-experiments through manipulation of the explanation increases external validity. It helps to isolate the effects of sociodemographic characteristics and perceived fairness on trust. This approach can be efficient in assessing the relationship between the variables and policy-relevant in terms of validity ecology Duarte et al., 2023.

3.1. Sample

The target population includes participants aged 18 and above from hard-to-reach groups in a specified region/country of interest referring to racially or ethnically diverse populations, Individuals Belonging to low-income families, disabled persons, and those who speak non-English languages. The selection of samples followed an accidental pro Ratio of race, income, and geographical area within the city or rural areas. The purposive sampling technique was used to select participants from various groups in the community, such as non-governmental organizations, advocacy groups, migrants, and Indigenous people. The participants samples are N = 240 with each stratum having 80 participants to conduct subgrouping. Two requirements of participants are that they must classify themselves as belonging to a disadvantaged group and that they have used AI or services that use AI at some point.

3.2. Data Collection

The questionnaires were administered online- via social media and social organizations- and face-to-face by community centers, using tablets or paper questionnaires for participants with limited access to digital facilities. Primary outcome measures were participant satisfaction with the respective AI policy and assessment outcomes, and secondary outcome measures will be perceived technical comprehension of the decision and acceptability of the AI policy. The demographic questions were asked at the end of each vignette, and participants were asked to complete validated scales.

3.3. Instrument Construction

The vignettes were developed during group discussions and meetings with community advisors implementing real-life issues like housing and unemployment. The explanations are of different levels of detail and cultural adaption and are provided in the local language while applying local cultural standards. The level of trust was established by administering the Trust in Automation Scale or TAS and an Algorithmic Distrust Scale. In contrast, perceived fairness was evaluated based on the procedural justice dimensions scale. Participants will also assess the adequacy of the explanations given by the normative theories to their everyday realities.

3.4. Pilot Testing

Vignettes and the respective scales were pre-tested with thirty participants from the target groups in a pilot study to ensure that the study's instrumentality was realistic. Cognitive interviews will seek to establish those areas that contain language that needs to be explained to participants in a lay manner to ensure that as many persons as possible can be sampled for the study.

3.5. Data Analysis

To analyze the results of trust, distrust, and fairness, H-Freeware 4.05 Categorical Data Package was used to get frequencies and means of each subgroup. Inferential statistical analysis included multiple linear regression to establish those variables that predicted the level of trust, excluding the variables of digital literacy, and MANOVA to compare trust results between the different vignettes offered. Cultural alignment will also be analyzed as a mediator between perceived fairness and top-level trust, where race and income are the moderators. Statistical analysis of the data was done using the statistical analysis software SPSS 28. For missing data, full information maximum likelihood was used. This is a highly effective approach for developing algorithmic accountability in as much as it carries the voices of the marginalized at the fag of the quantitative methodologies. The vignette experiments translate developed theoretical concepts of procedural justice into practice outcomes. Partnerships with communities boost ecological and cultural relevance.

4. Results

4.1. Frequency Distributions for Demographic Variables

The participants were assigned to the three experimental conditions in equal numbers. Table 01 shows that 33.3% were given the Technical Explanation, 33.3% the Plain-Language explanation, and 33.3% were exposed to the Human Decision condition. This balanced assignment allows for a proper comparison of conditions.

Table 1 Experimental Group Assignment (Across All Vignettes)

Explanation Type	Frequency	Percent	Valid Percent	Cumulative Percent
Technical Explanation	80	33.30%	33.30%	33.30%
Plain-Language	80	33.30%	33.30%	66.60%
Human Decision	80	33.30%	33.30%	100.00%
Total	240	100.00%	100.00%	

Table 02 shows that 37.5% of the respondents never used AI systems, which implies that only 62.5% of the respondents had previous exposure to the systems. This means that a majority met with some familiarity with AI and its application in public services; thus, they provide a good grounding view.

Table 2 Respondents Prior AI Exposure

Response	Frequency	Percent	Valid Percent	Cumulative Percent
Yes	150	62.50%	62.50%	62.50%
No	90	37.50%	37.50%	100.00%
Total	240	100.00%	100.00%	

The distribution of the education reveals that the respondents had various educational levels. Table 03 shows that 7.5% of them have less than a high school education, 32.5% have a high school diploma, 45.0% hold a bachelor's degree, and 15.0% have a graduate degree. It helps to understand perceptions of trust in AI and how educational background may affect its decision-making.

Table 3 Education Level of Respondents

Education Level	Frequency	Percent	Valid Percent	Cumulative Percent
No High School	20	8.30%	8.30%	8.30%
High School Diploma	80	33.30%	33.30%	41.60%
Bachelor's Degree	100	41.70%	41.70%	83.30%
Graduate Degree	40	16.70%	16.70%	100.00%
Total	240	100.00%	100.00%	

Table 04 shows that household income distribution shows moderate cross-sectional economic disparity. About 33.3% of the respondents earn below 20,000; 37.5% earn between 20,000 and 40,000; 18.8% earn between 40,000 and 60,000 and 10.4% earn more than 60,000. This can influence the public's perception of AI and determine whether they trust it, depending on their SES. In this case, the social status may affect the response behavior due to income differences.

Table 4 Household Income of Respondents

Income Category	Frequency	Percent	Valid Percent	Cumulative Percent
< 20k	80	33.30%	33.30%	33.30%
20k – 40k	90	37.50%	37.50%	70.80%
40k – 60k	45	18.80%	18.80%	89.60%
> 60k	25	10.40%	10.40%	100.00%
Total	240	100.00%	100.00%	

Table 05 shows that the ethnic distribution of the respondents is also spread across different regions. The largest ethnicity is Punjabi, with 50% of the population, followed by Sindhi, at 20.8%; Pashtun, at 16.7%; Balochi, at 8.3%; and Others, at 4.2%. This demographic variety is useful for understanding cultural factors related to trust in AI across subgroups of people. These, of course, are important for the implementation of policies that will suit the intended population.

Table 5 Ethnicity of Respondents

Ethnicity	Frequency	Percent	Valid Percent	Cumulative Percent
Punjabi	120	50.00%	50.00%	50.00%
Sindhi	50	20.80%	20.80%	70.80%
Pashtun	40	16.70%	16.70%	87.50%
Balochi	20	8.30%	8.30%	95.80%
Other	10	4.20%	4.20%	100.00%
Total	240	100.00%	100.00%	

4.2. Descriptive Statistics

Descriptive statistics included in Table 06 show that the variables of the survey had a moderate central tendency. The two variables of importance were perceived decision process trust, with a mean of 4.20 (SD = 1.55), and clarity of explanation, with an average of 4.40 (SD = 1.50). There was a moderate level of perceived fairness, with the mean score being 4.00 (1.60) for the study participants. Algorithmic distrust items produced a mean of 4.50 for bias (Standard deviation = 1.40) and 4.30 for powerlessness (Standard deviation = 1.60), which indicates moderate concern. The level of perceived fairness was slightly above average, with voice at 3.90 (SD = 1.70) and consistency at 4.00 (SD = 1.50) of 7. Cultural convergence reached 4.20 (respect) and 4.50 (language); digital competencies were relatively high, 4.60 (SD, 1.40). These findings show that respondents had a fairly favorable attitude towards each other's beliefs and opinions.

Table 6 Descriptive Statistics for Continuous Survey Measures

Variable	N	Mean	Std. Deviation	Minimum	Maximum
Trust in Decision Process	240	420.00%	155.00%	100.00%	7
Explanation Clarity	240	440.00%	150.00%	100.00%	7
Fair Treatment	240	400.00%	160.00%	100.00%	7
Algorithmic Distrust – Bias Item	240	450.00%	140.00%	100.00%	7
Algorithmic Distrust – Powerlessness	240	430.00%	160.00%	100.00%	7
Perceived Fairness – Voice Item	240	390.00%	170.00%	1	7
Perceived Fairness – Consistency	240	4	1.5	1	7
Cultural Alignment – Respect	240	4.2	1.4	1	7
Cultural Alignment – Language	240	4.5	1.3	1	7
Digital Literacy	240	4.6	1.4	1	7

4.3. Effect of Explanation Type on Trust in Decision Process (One-Way ANOVA)

Table 07 shows that explanation type was a significant factor in influencing trust in the decision process, F (2,237) = 10.5, p = .01. The post hoc analysis revealed that the Technical Explanation group had a significantly lower level of trust than that of the Control (t = 4.54, p < .001) and Plain-Language (t = -1.98, p = .048) groups. The comparison of the two groups, Plain-Language and Control, was not very different (p < .05 = .095). This implies that human decisions are preferred over automation, but simpler explanations from such automation systems are considered better.

Table 7 Effect of Explanation Type on Trust

One Way Anova

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	38.5	2	19.25	10.5	0.01
Within Groups	432	237	1.82		
Total	470.5	239			

Post Hoc Tests (Bonferroni)
Technical vs Plain-Language: p = .048
Technical vs Control: p < .001
Plain-Language vs Control: p = .095

4.4. Effect of Explanation Type on Combined Dependent Variables (MANOVA)

Table 08 shows that the explanation type has a multivariate effect with the values of trust, fairness, and explanation clarity, Pillai's Trace = .15, F (6,474) = 3.50, p = .002. Consequently, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root revealed significance. Thus, the results show that the explanation style does impact several dependent variables, and human decisions are more trusted than plain-language AI explanations.

Table 8 Effect of Explanation Type

Multivariate Tests

Multivariate Tests

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.
Explanation Type	0.15	3.5	6	474	0.002

Other Test Statistics

Test Statistic	Value	F Value	p-value
Wilks' Lambda	0.85	F (6,474) = 3.50	0.002
Hotelling's Trace	0.176	F (6,474) = 3.51	0.002
Roy's Largest Root	0.15	F(2,238) = 8.25	0

4.5. Cultural Alignment on Perceived Fairness – Regression

The regression results presented in Table 09 show that cultural affiliation is a significant predictor of perceived fairness; B = 0.50, t(238) = 5.00, p < .001. The model accounted for 20 % of the variance in the criteria (R^2 = .20, p < .001). The results indicated that when participants had a high perception of the fairness of the AI, this was linked to a high cultural congruity, meaning that the people were convinced that the decision-making process of an AI system was culturally acceptable.

Table 9 Effect of Cultural Alignment on Perceived Fairness

Model 01- Outcome: Cultura Alignment

Model	R	R Square	Adjusted R Square	Std. Error
1	0.447	0.2	0.195	1.2

ANOVA

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	150	1	150	25	0
Residual	600	238	2.52		
Total	750	239			

Coefficient Summary

Predictor	Unstandardized Coeff. (B)	Std. Error	t	Sig.
(Constant)	2.5	0.4	6.25	0
Perceived Fairness	0.5	0.1	5	0

4.6. Trust in the Decision Process on Perceived Fairness and Cultural

Table 10 shows that both perceived fairness and cultural alignment have a significant influence on the participants' level of trust in decisions made by AI with the coefficients and p-values of B = .30, p = .013, and B = .40, p < .001, respectively. This regression model accounted for 30% of the variance in the data (R^2 = .30, F (6, 92) = 10.63, p < .001). The results proved structural congruence fully mediates the relationship between fairness and trust, with an effect of 95% CI 0.20(0.10, 0.35).

Table 10 Effect of Trust in Decision Process

Model 2 – Outcome: Trust in Decision Process

Model	R	R Square	Adjusted R Square	Std. Error
1	0.547	0.3	0.29	1.1

Variance in Model

Source	Sum of Squares	df	Mean Square	F	Sig.
Regression	180	2	90	22.5	0.03
Residual	420	237	1.77		
Total	600	239			

Coefficient

Predictor	Unstandardized Coeff. (B)	Std. Error	Standardized Beta	T	Sig.
(Constant)	1.8	0.5		3.6	0
Perceived Fairness	0.3	0.12	0.25	2.5	0.013
Cultural Alignment	0.4	0.08	0.45	5	0

Indirect Effect

Indirect Effect	Calculation	Value
Indirect Effect (ab)	(0.50) × (0.40)	0.2
Bootstrapped 95% CI	[0.10, 0.35]	

Bootstrapping Analysis (5,000 samples)

Effect	Bootstrapped Estimate	SE	Lower 95% CI	Upper 95% CI
Indirect Effect (ab)	0.2	0.07	0.1	0.35

4.7. Moderation of the Indirect Effect by Income Level on the X→M Path

Table 11 highlights the mediating role of income level on cultural alignment, and the indirect relationship between fairness and trust was also found to be more significant in low-income consumers (ß = 0.25, CI = 0.10-0.45) as compared to high-income consumers (ß = 0.15, CI = -0.02-0.32). This implies that lower-income earners are more sensitive to perceived fairness, thus advocating for culturally appropriate AI policies in poor areas.

Table 11 Effect of Income Level on Dependent Variables

Income Level	Indirect Effect	Boot SE	Lower 95% CI	Upper 95% CI
Low (<20k)	0.25	0.08	0.1	0.45
Medium (20k-40k)	0.2	0.07	0.08	0.38
High (>40k)	0.15	0.09	-0.02	0.32

5. Discussion

The present study examines trust and distrust in Explainable AI (XAI) among marginalized groups. The study focuses on demographic factors and their perception of fairness; it assumes a direct relationship between them and trust. This might obscure the existence of interactional effects or contextual effects, which the literature on procedural justice has shown can be existent and distinguishable (Jiang et al., 2023; Açıkgöz et al., 2020). A cross-sectional survey method was employed with embedded mock scenarios, which includes sensible yet has limitations of a cross-sectional design and cannot cover the temporal changes in trust perceptions (Orabi et al., 2024). While purposive sampling improves representation, it brings in self-selection bias, which may, in turn, affect the extent of generalization of studies (Cunha, 2024). The integration of qualitative methodological components could add more depth to the interpretation and increase the understanding of the multifaceted nature of trust processes (Schaap & Saarikkomäki, 2022).

The study findings show that the type of explanation given matters greatly, with human decisions holding the most trust, plain language options being the next most trusted, and, lastly, technical options. This observation is consistent with the PJT since the decision-making process was clear, and participants perceived it as fair (Lowrey-Kinberg et al., 2020; Deb et al., 2023). However, it was found that while simple explanations indeed help in improving the degree of trust in a general manner, this concept may not apply in all settings. For instance, people with formal education or with computer literacy may regard simplistic explanations as opaque, giving the impression that there is a direct linkage between the kinds of explanations offered and the level of trust observed, which is not the case (Buccella, 2022).

Further, it is helpful to identify cultural alignment as a factor in perceived fairness, but it is unclear how one operationalizes cultural alignment. The standards of perceived credibility, therefore, differ from one culture to another, making it difficult to set a standard when it comes to the trustworthiness of an AI (Yang & Lee, 2024; Niklas & Dencik, 2024). Further research should explore whether AI systems with cultural awareness create long-term trust or whether, in fact, they result in trust that only lasts for the duration of the interaction since the latter may indicate a misunderstanding of trust-related processes (Aguilar-Rojas et al., 2024).

The regression analysis findings emphasized that perceived fairness partially explains trust in AI, although fairness is a concept whose definition is not universally agreed on. This means that different groups will have different perceptions of what is right or fair, thus posing challenges to AI governance (Hermansyah et al., 2023; Ritchie et al., 2021). Future studies could expand these two aspects more specifically in the discussion of the algorithm's prejudice, including where technical methods of fairness, such as a rate of equal errors, are not necessarily the same as the social methods, for instance, addressing past injustice (Hong & Chen, 2024). Low income is found to increase the importance of fairness perception; thus, economic vulnerability has been established to give the test of trust. However, the study fails to establish whether the distrust stems from the AI systems or the skepticism that people have for public institutions (Robert et al., 2020).

In addition, MANOVA results prove that explanation type does affect trust, fairness, and clarity, but low R-squared indices indicate that other variables mediate the phenomenon. Other potential variables are risk-taking propensity, previous experiences in interacting with AI, and level of trust in institutions, which could also influence how credible AI is considered by respondents (Yurnalis & Mangundjaya, 2020; Bankins et al., 2022). Also, it noted that respondents who had prior exposure to AI still expressed similar distrust. Thus, it would also be important to make one understand that exposure to AI simply meant that one needed to trust the design of explainability had to be made more inclusive and transparent (Talukder & Shompa, 2024).

In general, this study significantly contributes to the existing body of knowledge on AI governance by shedding light on the relationship between explainability, fairness, and trust to marginalized populations. It provides suggestions concerning public sector AI policies and employs culturally sensitive and explainable decision-making paradigms (Krarup & Horst, 2023; Hsieh et al., 2024). However, building trust in AI is not only a matter of technological development but requires organizational responsibilities, people's involvement, and cooperation (Morin-Martel, 2023; Khan & Mishra, 2023). The current policymakers need to understand that algorithm fairness is not a mathematical problem to solve but also a socio-political process that has to be constantly monitored and discussed with the concerned community (Zhu & Park, 2022; Vargas-Murillo et al., 2024). Although the study provides the quantitative background for future works, it is advisable to include the qualitative evaluation to reveal the additional aspects of the trust worrying. Cross-sectional studies with longitudinal evaluation and additional experimental intercessions might augment the relevance of the outcomes in real-world settings (Asante et al., 2024; Trang et al., 2024).

6. Conclusion

This study focuses on the three major aspects of trust in artificial intelligence for policy-making, including explainability, fairness, and cultural relevance for the latter. The study findings show that oversimplified and overexplained information increases confidence, but human decisions are most trusted. In addition, perceived fairness moderates the relationship between trust and the degree to which an individual is engaged, with the lower-income groups being especially sensitive to the fairness perceptions, highlighting the socio-economic aspect of algorithmic decision-making. However, the findings showed that Trust, cannot be a purely technical concern that will be fixed by making AI systems more transparent and it needs institutional response, inclusion, and social consciousness. More research should be conducted on the long-term changes in trust, psychological factors, and qualitative studies. Thus, it is crucial to pay attention to AI systems being pseudo using cultural values and expectations in developing effective policies to guard the equitable and ethical decision-making process.

References

Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Machine learning approaches for predicting maize biomass yield: leveraging feature engineering and comprehensive data integration. Sustainability, 17(1), 256. https://doi.org/10.3390/su17010256

Açıkgöz, Y., Davison, H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. International Journal of Selection and Assessment, 28(4), 399-416. https://doi.org/10.1111/ijsa.12306

Aguilar-Rojas, Ó., Herrera, C., & Pérez-Rueda, A. (2024). The importance of social comparison in perceived justice during the service recovery process. European Journal of Management and Business Economics, 33(4), 488-504. https://doi.org/10.1108/ejmbe-02-2023-0056

Amarasinghe, K., Rodolfa, K., Lamba, H., & Ghani, R. (2023). Explainable machine learning foEU's ac policy: use cases, gaps, and research directions. Data & Policy, 5. https://doi.org/10.1017/dap.2023.2

Andersen, J., Nota, P., Boychuk, E., Schimmack, U., & Collins, P. (2023). Reducing racial bias and lethal force among Canadian police officers: concrete recommendations for change. Canadian Journal of Behavioural Science/ Revue Canadienne Des Sciences Du Comportement, 55(2), 153-160. https://doi.org/10.1037/cbs0000364

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

Arrieta, A., Díaz-Rodríguez, N., Ser, J., Bennetot, A., Tabik, S., Barbado, A., … & Herrera, F. (2020). Explainable artificial intelligence (xai): concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Asante, K., Sarpong, D., & Boakye, D. (2024). On the consequences of a bias: when moral values supersede algorithm bias. Journal of Managerial Psychology. https://doi.org/10.1108/jmp-05-2024-0379

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). Ai decision-making with dignity? Contrasting workers' justice perceptions of human and AI decision-making in a human resource management context. Information Systems Frontiers, 24(3), 857-875. https://doi.org/10.1007/s10796-021-10223-8

Barajas-Gonzalez, R., Ayón, C., Brabeck, K., Rojas-Flores, L., & Valdez, C. (2021). An ecological expansion of the adverse childhood experiences (aces) framework to include threat and deprivation associated with U.S. immigration policies and enforcement practices: an examination of the Latino immigrant experience. Social Science & Medicine, 282, 114126. https://doi.org/10.1016/j.socscimed.2021.114126

Baykurt, B. (2022). Algorithmic accountability in U.S. cities: transparency, impact, and political economy. Big Data & Society, 9(2). https://doi.org/10.1177/20539517221115426

Bell, A., Nov, O., & Stoyanovich, J. (2023). Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. Data & Policy, 5. https://doi.org/10.1017/dap.2023.8

Birhane, A. (2022). Power to the people? Opportunities and challenges for participatory AI.. https://doi.org/10.48550/arxiv.2209.07572

Buccella, A. (2022). "ai for all" is a matter of social justice. Ai and Ethics, 3(4), 1143–1152. https://doi.org/10.1007/s43681-022-00222-z

Cech, F. (2021). The agency of the forum: mechansms for algorithmic accountability through the lens of the agency. Journal of Responsible Technology, 7–8, 100015. https://doi.org/10.1016/j.jrt.2021.100015

Christou, I., Soldatos, J., Papadakis, T., Gutiérrez-Rojas, D., & Nardelli, P. (2023). Feature selection via minimal covering sets for industrial Internet of Things applications.. https://doi.org/10.1109/dcoss-iot58021.2023.00092

Clarke, D., Sakata, S., & Barahona, S. (2023). Public policy uses of the SEEA stocks and flows accounts.

Coyle, D. and Weller, A. (2020). "explaining" machine learning reveals policy challenges. Science, 368(6498), 1433–1434. https://doi.org/10.1126/science.aba9647

Cunha, C. (2024). An exploratory study on the practice of procedural justice and use of force in police-citizen encounters.. https://doi.org/10.21428/88de04a1.6c1ae65b

Deb, S., Nafi, S., Mallik, N., & Valeri, M. (2023). Mediating effect of emotional intelligence on the relationship between employee job satisfaction and firm performance of small business. European Business Review, 35(5), 624-651. https://doi.org/10.1108/ebr-12-2022-0249

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in AI design: theoretical foundations and the current state of practice., 1-23. https://doi.org/10.1145/3617694.3623261

Duarte, R., Correia, F., Arriaga, P., & Paiva, A. (2023). AI trust: Can explainable AI enhance warranted trust? Human Behavior and Emerging Technologies, 2023, 1-12. https://doi.org/10.1155/2023/4637678

Ekstrom, P., Forestier, J., & Lai, C. (2021). Racial demographics explain the link between racial disparities in traffic stops and county-level racial attitudes.. https://doi.org/10.31234/osf.io/znhuf.

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Medical Informatics and Decision Making, 20(1). https://doi.org/10.1186/s12911-020-01191-1

González, R. (2024). Algorithmic policing: Part 1. Tech startups, venture capital, and law enforcement in America. Anthropology Today, 40(5), 23–27. https://doi.org/10.1111/1467-8322.12916

Guo, Y., Yang, X., & Shi, C. (2023). Enabling team of teams: a trust inference and propagation (tip) model in multi-human multi-robot teams.. https://doi.org/10.15607/rss.2023.xix.003

Gursoy, F. and Kakadiaris, I. (2022). System cards for ai-based decision-making for public policy.. https://doi.org/10.48550/arxiv.2203.04754

Hermansyah, M., Najib, A., Farida, A., Sacipto, R., & Rintyarna, B. (2023). Artificial intelligence and ethics: building an artificial intelligence system that ensures privacy and social justice. International Journal of Science and Society, 5(1), 154-168. https://doi.org/10.54783/ijsoc.v5i1.644

Hong, W. and Chen, C. (2024). Ethical concerns upon artificial intelligence empowered human resource management: a qualitative study among middle-level managers from Beijing technology companies. International Journal for Multidisciplinary Research, 6(5). https://doi.org/10.36948/ijfmr.2024.v06i05.28860

Hsieh, C., Chen, S., Peng, T., Chen, P., Chen, A., & Chen, C. (2024). The relationship between workplace justice and self-evaluated nonfatal occupational accidents among healthcare employees in Taiwan: an observational study. Medicine, 103(32), e39215. https://doi.org/10.1097/md.0000000000039215

Hung, T. and Yen, C. (2023). Predictive policing and algorithmic fairness. Synthese, 201(6). https://doi.org/10.1007/s11229-023-04189-0

Jeong, N., Epsztein, R., Wang, R., Park, S., Lin, S., & Tong, T. (2023). Exploring the knowledge attained by machine learning on ion transport across polyamide membranes using explainable artificial intelligence. Environmental Science & Technology, 57(46), 17851-17862. https://doi.org/10.1021/acs.est.2c08384

Jiang, L., Qin, X., Yam, K., Dong, X., Liao, W., & Chen, C. (2023). Who should be first? How and when ai-human order influences procedural justice in a multistage decision-making process. Plos One, 18(7), e0284840. https://doi.org/10.1371/journal.pone.0284840
Jimenez, T., Helm, P., & Arndt, J. (2022). Racial prejudice predicts police militarization. Psychological Science, 33(12), 2009-2026. https://doi.org/10.1177/09567976221112936

Kaushik, P. (2020). Impact and usage of AI in the public sector. International Journal of Engineering in Computer Science, 2(1), 38–43. https://doi.org/10.33545/26633582.2020.v2.i1a.99

Khan, A. and Mishra, A. (2023). Ai credibility and consumer-ai experiences: a conceptual framework. Journal of Service Theory and Practice, 34(1), 66–97. https://doi.org/10.1108/jstp-03-2023-0108

Khan, M. & Vice, J. (2022). Toward accountable and explainable artificial intelligence part two: the framework implementation.. https://doi.org/10.36227/techrxiv.19102094

Knox, D., Lowe, W., & Mummolo, J. (2020). Administrative records mask racially biased policing. American Political Science Review, 114(3), 619-637. https://doi.org/10.1017/s0003055420000039

Kochel, T. & Nouri, S. (2024). Impact of in-service implicit bias training: a study of attitudinal changes and intention to apply anti-bias techniques. Police Quarterly, 27(4), 561–581. https://doi.org/10.1177/10986111241237512

Krarup, T. and Horst, M. (2023). European artificial intelligence policy as digital single market making. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231153811

Lai, C. and Lisnek, J. (2023). The impact of implicit-bias-oriented diversity training on police officers' beliefs, motivations, and actions. Psychological Science, 34(4), 424-434. https://doi.org/10.1177/09567976221150617

Laniyonu, A. and Donahue, S. (2023). Effect of racial misclassification in police data on estimates of racial disparities. Criminology, 61(2), 295-315. https://doi.org/10.1111/1745-9125.12329

Liao, Q., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences., 1-15. https://doi.org/10.1145/3313831.3376590

Lowrey-Kinberg, B., Mellinger, H., & Kearns, E. (2020). How social dominance orientation shapes perceptions of police. Policing an International Journal, 43(4), 607–624. https://doi.org/10.1108/pijpsm-02-2020-0022

Misra, S., Das, S., Gupta, S., & Sharma, S. (2020). Public policy and regulatory challenges of artificial intelligence (ai)., 100–111. https://doi.org/10.1007/978-3-030-64849-7_10

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., … & Becker, B. (2023). Trust toward humans and trust toward artificial intelligence are not associated: initial insights from self-report and neurostructural brain imaging. Personality Neuroscience, 6. https://doi.org/10.1017/pen.2022.5

Morin-Martel, A. (2023). Machine learning in bail decisions and judges' trustworthiness. Ai & Society, 39(4), 2033-2044. https://doi.org/10.1007/s00146-023-01673-6

Mucci, A., Green, W., & Hill, L. (2024). Incorporation of artificial intelligence in healthcare professions and patient education to foster effective patient care. New Directions for Adult and Continuing Education, 2024(181), 51-62. https://doi.org/10.1002/ace.20521

Niklas, J. and Dencik, L. (2024). Data justice in the "twin objective" of market and risk: how discrimination is formulated in EU's policy. Policy & Internet, 16(3), 509-522. https://doi.org/10.1002/poi3.392

Njalsson, G. (2023). Toward an integrated model for public technology policy analysis - a taxonomy useful for determining scope and type of analysis. Rudn Journal of Public Administration, 10(2), 269–285. https://doi.org/10.22363/2312-8313-2023-10-2-269-285

Orabi, T., Alahewat, A., Abualfalayeh, G., & Samara, H. (2024). The interdisciplinary nature of AI and human resource management: a bibliometric analysis. Human Systems Management, 43(6), 845-871. https://doi.org/10.3233/hsm-240054

Papadakis, T., Christou, I., Ipektsidis, C., Soldatos, J., & Amicone, A. (2024). Explainable and transparent artificial intelligence for public policymaking. Data & Policy, 6. https://doi.org/10.1017/dap.2024.3

Payne, B. & Rucker, J. (2022). Explaining the spatial patterning of racial disparities in traffic stops requires a structural perspective: further reflections on Stelter et al. (2022) and Ekstrom et al. (2022). Psychological Science, 33(4), 666–668. https://doi.org/10.1177/09567976211056641

Ploug, T., Sundby, A., Moeslund, T., & Holm, S. (2021). Population preferences for performance and explainability of artificial intelligence in health care: choice-based conjoint survey. Journal of Medical Internet Research, 23(12), e26611. https://doi.org/10.2196/26611

Purves, D. & Davis, J. (2022). Public trust, institutional legitimacy, and the use of algorithms in criminal justice. Public Affairs Quarterly, 36(2), 136–162. https://doi.org/10.5406/21520542.36.2.03

Rios, V., Prieto, G., & Ibarra, J. (2020). Mano suave–mano dura: legitimacy policing and Latino stop-and-frisk. American Sociological Review, 85(1), 58-75. https://doi.org/10.1177/0003122419897348

Ritchie, K., Cartledge, C., Growns, B., Yan, A., Wang, Y., Guo, K., … & White, D. (2021). Public attitudes towards the use of automatic facial recognition technology in criminal justice systems around the world. Plos One, 16(10), e0258241. https://doi.org/10.1371/journal.pone.0258241

Robert, L., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. Human-Computer Interaction, 35(5-6), 545-575. https://doi.org/10.1080/07370024.2020.1735391

Robles, P. & Mallinson, D. (2023). Artificial intelligence technology, public trust, and effective governance. Review of Policy Research, 42(1), 11–28. https://doi.org/10.1111/ropr.12555

Rojas, E. and Li, M. (2024). Trust is contagious: social influences in the human-human-ai team. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68(1), 317-322. https://doi.org/10.1177/10711813241262025

Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: a systematic review. Jmir Ai, 3, e53207. https://doi.org/10.2196/53207

Schaap, D. and Saarikkomäki, E. (2022). Rethinking police procedural justice. Theoretical Criminology, 26(3), 416-433. https://doi.org/10.1177/13624806211056680

Schiff, D. (2023). Looking through a policy window with tinted glasses: setting the agenda for U.S. AI policy. Review of Policy Research, 40(5), 729–756. https://doi.org/10.1111/ropr.12535

Schmid, N., Haelg, L., Sewerin, S., Schmidt, T., & Simmen, I. (2020). Governing complex societal problems: the impact of private on public regulation through technological change. Regulation & Governance, 15(3), 840-855. https://doi.org/10.1111/rego.12314

Susser, D. (2021). 12. Predictive policing and the ethics of preemption. 268–292. https://doi.org/10.18574/nyu/9781479803729.003.0013

Talukder, K. and Shompa, T. (2024). Artificial intelligence in criminal justice management: a systematic literature review. NHJ, 1(01), 63-82. https://doi.org/10.70008/jmldeds.v1i01.42

Toll, D., Lindgren, I., Melin, U., & Madsen, C. (2020). Values, benefits, considerations, and risks of AI in government. Jedem - Ejournal of Edemocracy and Open Government, 12(1), 40-60. https://doi.org/10.29379/jedem.v12i1.593

Tran, Y. & Nguyen, N. (2020). The impact of the performance measurement system on the organizational performance of the public sector in a transition economy: Is public accountability a missing link? Cogent Business & Management, 7(1), 1792669. https://doi.org/10.1080/23311975.2020.1792669

Trang, N., Linh, N., Hoang, N., Kiet, P., Loan, L., & Phuc, N. (2024). Right to a fair trial when applying artificial intelligence in criminal justice - lessons and experiences for Vietnam. Journal of Law and Sustainable Development, 12(3), e601. https://doi.org/10.55908/sdgs.v12i3.601

Vargas-Murillo, A., Pari-Bedoya, I., Turriate-Guzmán, A., Delgado-Chávez, C., & Sanchez-Paucar, F. (2024). Transforming justice: implications of artificial intelligence in legal systems. Academic Journal of Interdisciplinary Studies, 13(2), 433. https://doi.org/10.36941/ajis-2024-0059

Veer, S., Riste, L., Cheraghi-Sohi, S., Phipps, D., Tully, M., Bozentko, K., … & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. Journal of the American Medical Informatics Association, 28(10), 2128-2138. https://doi.org/10.1093/jamia/ocab127

Wang, X., Lu, Z., & Yin, M. (2022). Will you accept the AI recommendation? Predicting human behavior in ai-assisted decision making., 1697-1708. https://doi.org/10.1145/3485447.3512240

Wang, Y., Chen, Y., Chien, S., & Wang, P. (2024). Citizens' trust in ai-enabled government systems. Information Polity, 29(3), 293-312. https://doi.org/10.3233/ip-230065

Wei, S., Ge, Z., & Tan, C. (2024). The public interest in the digital age: exploring the emerging roles and governance models of the AI as a common good. Journal of Ecohumanism, 3(7), 2529-2544. https://doi.org/10.62754/joe.v3i7.4397

Weng, F., Cheng, D., Zhuang, M., Lü, X., & Yang, C. (2024). The effects of governance quality on renewable and nonrenewable energy consumption: an explainable decision frame. Journal of Forecasting, 43(6), 2146-2162. https://doi.org/10.1002/for.3110

Yan, L., Diao, Y., Lang, Z., & Gao, K. (2020). Corrosion rate prediction and influencing factors evaluation of low-alloy steels in the marine atmosphere using machine learning approach. Science and Technology of Advanced

Materials, 21(1), 359-370. https://doi.org/10.1080/14686996.2020.1746196

Yang, Q. and Lee, Y. (2024). Ethical AI in financial inclusion: the role of algorithmic fairness on user satisfaction and recommendation.. https://doi.org/10.20944/preprints202407.1655.v1

Young, M., Ehsan, U., Singh, R., Tafesse, E., Gilman, M., Harrington, C., … & Metcalf, J. (2024). Participation versus scale: Tensions in the practical demands on participatory AI. First Monday. https://doi.org/10.5210/fm.v29i4.13642

Yurnalis, Y. and Mangundjaya, W. (2020). Testing the impact of organizational justice on affective commitment to change with work engagement as mediator.. https://doi.org/10.2991/assehr.k.200407.018

Zhang, B. & Dafoe, A. (2020). U.S. public opinion on the governance of artificial intelligence., 187–193. https://doi.org/10.1145/3375627.3375827

Zhu, T. & Park, S. (2022). Encouraging brand evangelism through failure attribution and recovery justice: the moderating role of emotional attachment. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.877446

Ziosi, M. and Pruss, D. (2024). Evidence of what, for whom? The socially contested role of algorithmic bias in a predictive policing tool., 1596-1608. https://doi.org/10.1145/3630106.3658991