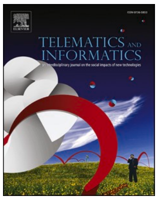
Telematics and Informatics 91 (2024) 102135

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele



EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives

Ching-Hua Chuan [a], Ruoyu Sun [b], Shiyun Tian [c], Wan-Hsiu Sunny Tsai [d,*]

[a] Department of Interactive Media, University of Miami, 5100 Brunson Drive, Coral Gables, FL 33134, USA
[b] Department of Advertising and Public Relations, University of Georgia, 120 Hooper St, Athens, GA 30602-3018, USA
[c] Department of Marketing, Sacred Heart University, 5151 Park Avenue, Fairfield, CT 06825, USA
[d] Department of Strategic Communication, University of Miami, 5100 Brunson Drive, Coral Gables, FL 33134, USA

ARTICLE INFO

Keywords:
eXplainable artificial intelligence (XAI)
Algorithmic bias
Fairness
Inclusiveness
Imposed users

ABSTRACT

This study explored the potential of eXplainable Artificial Intelligence (XAI) in raising user awareness of algorithmic bias. This study examined the popular "explanation by example" XAI approach, where users receive explanatory examples resembling their input. As this XAI approach allows users to gauge the congruence between these examples and their circumstances, perceived incongruence then evokes perceptions of unfairness and exclusion, prompting users not to put blind trust in the system and raising awareness of algorithmic bias stemming from non-inclusive datasets. The results further highlight the moderating role of users' prior experience with discrimination.

1. Introduction

Artificial intelligence (AI) has been widely recognized as a double-edged sword, improving every aspect of our lives and at the same time posing critical threats to societal well-being, particularly because AI is rapidly being integrated into decision-making processes, from resume screening for employment decisions to financial loan approval. However, AI biases based on race, gender, sexual orientation, and other identity factors have been increasingly observed in the outputs of AI systems (Kordzadeh and Ghasemaghaei, 2022). Algorithmic bias can be found in common, everyday applications, such as search engines' autocomplete predictions that produce more negative biases for female, Black, and homosexuality-related prefixes (Lin et al., 2023), and in domain-specific tools, such as sentencing and risk assessment software in criminal justice that discriminates against Black defendants (Angwin et al., 2016). As AI algorithms are deployed via industry-wide mass automation, disparities and inequities are exacerbated at a much greater scale and speed, generating growing concerns on algorithmic bias (Chuan et al., 2023).

Algorithmic bias is defined as "systematic and structured errors in an artificial intelligence system that generate unfair results and inequalities" (Shin and Shin, 2023, p. 90). Critically, due to the "black-box" nature of complicated AI systems, even AI developers cannot reliably explain why a specific (potentially biased) decision is produced. Therefore, facilitating acute awareness and recognition of potential algorithmic biases among various stakeholders within the AI ecosystem is a crucial first step for mitigating the

* Corresponding author.
E-mail addresses: c.chuan@miami.edu (C.-H. Chuan), rsun@uga.edu (R. Sun), tians@sacredheart.edu (S. Tian), wanhsiu@miami.edu (W.-H.S. Tsai).

https://doi.org/10.1016/j.tele.2024.102135
Received 8 February 2024; Received in revised form 29 April 2024; Accepted 9 May 2024
Available online 10 May 2024

adverse impacts of such biases and informing fair and sound decision-making.

To address algorithmic bias, a set of fairness, accountability, transparency, and ethics (FATE) principles are now being demanded by scholars for decision-supporting AI systems. Focusing on the transparency principle, scholars in the research domain of eXplainable AI (XAI) have strived to transform the AI black box into a "glass box" (Rai, 2020) by providing technical solutions that automatically generate explanations on how a particular decision or recommendation was generated instead of another. The importance for XAI is underscored by the European Union's (EU's) General Data Protection Regulation (GDPR) that explicitly emphasizes the "right to explanation" as a person who is a data subject has the right to "an explanation of the decision reached after [algorithmic] assessment" (Selbst and Powles, 2018). In other words, companies can no longer use AI systems to make certain decisions about people living in the EU, unless the company can explain the decision. Moreover, scholars argue that model explanations can be used for algorithmic auditing. These explanations will make biases and unfairness more visible and easier to detect by revealing what the system is doing and how it arrives at its decisions (Kenny et al., 2021; Rong et al., 2023; Saeed and Omlin, 2023).

As explainability is inherently a human-centric consideration related to perception and cognition (Rong et al., 2023), user studies are an impetus for propelling the field of XAI. While scholars have increasingly advocated for a human-centered approach (e.g., Kong et al., 2024), XAI had been touted as a technical panacea primarily studied in computer science and relying on functionality-grounded evaluations that can be computed without human subjects (Nauta et al., 2023). In fact, only roughly 20 % of XAI evaluation studies involved human participants (Nauta et al., 2023), and social science theories regarding human information processing and decision making were rarely incorporated (Rong et al., 2023; Zhou et al., 2020). Consequently, most research has failed to meaningfully contextualize XAI within theoretical frameworks that are critical for illuminating the psychological mechanism underlying the effects of XAI. In a similar vein, when human perspectives were incorporated, early XAI research had primarily focused on experts (e.g., AI developers) and professional end users (e.g., medical doctors leveraging AI tools in diagnosis). Such perspectives ignore the needs of non-expert users, leaving out those who are most impacted by AI decisions and who have legitimate concerns about unfair treatment by the AI system (Kong et al., 2024). Therefore, this study aims to broaden the growing stream of user studies in XAI by focusing on non-expert users.

The limited user studies in XAI have predominantly focused on interpretability, defined as the system's ability to explain or present model predictions in terms people can understand (Rong et al., 2023). Consequently, common evaluation metrics include user understanding of the AI model, perceived explanation usefulness, and associated considerations of user satisfaction and trust (e.g., Mohseni et al., 2021). Empirical evidence on the potential of XAI for enhancing bias awareness and detection is scarce. It is particularly alarming that preliminary research findings suggest that XAI may persuade users to view AI errors as more reasonable and even correct (Kenny et al., 2021). In other words, the explanatory power of XAI may induce users' blind trust in AI systems, convincing them to follow advice even when it is incorrect (Van der Waa et al., 2021). Such blind trust undermines the human oversight that is imperative for mitigating the negative impacts of algorithmic biases.

By focusing on XAI, this study also aims to catalyze the growing interdisciplinary research on fairness in AI. Among the four principles of responsible AI, fairness is one of the most widely studied. The literature on algorithmic fairness has emphasized the avoidance and prevention of discriminatory or inequitable outcomes (Shin and Park, 2019). In particular, biased training data have been consistently cited as a cause of unfair AI decisions (Islam et al., 2022; Zhou et al., 2020). Various technical approaches have been proposed to ensure fairness in the development of AI decision-making systems by computer scientists and statisticians. However, achieving fairness in AI algorithms is more than a technical challenge—it also requires a deeper understanding of how users perceive and interpret fairness (Starke et al., 2022). In a series of algorithm user studies, Shin and colleagues provided important empirical evidence on the crucial role of users' fairness assessment in driving key outcomes, including algorithm acceptance, adoption, satisfaction, and trust (Shin and Park, 2019; Shin, 2020b; 2021a; Shin et al., 2020). Dodge et al. (2019) suggest that XAI can provide an interface for humans in the loop, enabling users to identify and address fairness issues in AI systems. This study thus expands user studies on fairness in AI to offer much needed insights not only for developers of AI systems, but also for the companies adopting these systems and policymakers shaping AI regulations.

To illuminate the social impacts of XAI as an emerging technology, this study presents one of the earliest empirical efforts to examine whether the popular XAI approach of "explanation by example" helps users identify lack of diversity in AI training datasets–one of the most frequently attributed causes of algorithmic bias (Hooker, 2021). Via an online experiment informed by fairness heuristic theory and social categorization theory, this study illuminates the mechanism underlying the effects of the post-hoc, personalized explanation on users' trust in the system and their recognition of algorithmic bias, through perceptions of fairness and inclusiveness. Research suggests that individuals vary in their sensitivity to structural biases and inequalities (Curtin et al., 2015; Kim, 2013). As experiencing discrimination often predisposes the impacted individuals to be more mindful and cognizant of injustice (Tanghe, 2016), we also incorporated users' prior experience with discrimination as an individual difference variable to examine its role in moderating XAI's effects on algorithmic bias detection.

1.1. Public awareness of algorithmic bias

As Shin (2023b) astutely pointed out, algorithms should be understood and studied as sociotechnical systems that reflect and reproduce social dynamics, including prejudices and inequalities. The danger of algorithmic bias has been widely discussed. For instance, media platforms powered by biased algorithms can exacerbate fake news (Giansiracusa, 2021; Shin and Shin, 2023). Algorithmic bias can be caused by problematic human decisions. For example, a nationally deployed healthcare algorithm in the United States routinely enrolled White patients into intensive-care programs ahead of Black patients who were sicker because medical cost data from insurance claims was used as a proxy for healthcare needs (Obermeyer et al., 2019). Additionally, as most contemporary

AI and machine learning techniques learn from historical data, they then embed inherited biases and disparities into the data. Biased data can be those that lack diversity or representations of different populations. For example, the disproportional employment rates on gender in the tech industries result in algorithms that prefer male applicants over female by penalizing applicants' resumes that include feminine keywords (e.g., women's chess club) (Dastin, 2022).

There exists limited research on what factors could drive individuals' perception and response toward algorithmic bias via experiments and interviews. Parra et al. (2021) studied participants' reactions to a biased outcome from an algorithm, i.e., whether they would question the outcome, in different everyday life scenarios. The biased outcome was presented as the participant and their friend who share all the identifiable personal factors (e.g., education level, socioeconomic achievement) except a key identity aspect (e.g., race or gender) but receive different outcomes from the algorithm. They found that the participants are more likely to question the algorithm in human resources recruitment and financial products scenarios, but least likely to do so in the healthcare scenario. Moreover, their study participants are more likely to question the algorithm in racial bias conditions than gender bias. However, research suggests that users' attitude toward potential algorithmic bias depends on whether they may benefit from the biased results. That is, personal gains from a biased algorithm may lead users to deny or ignore the bias against other demographic groups (Eslami et al., 2019; Wang et al., 2020).

Focusing on race-related algorithmic bias that tends to generate greater attention and discourse than gender bias (e.g., Parra et al., 2021), this study evaluates facial recognition algorithms that have been known to perform most accurately for lighter-skinned users but poorly for darker-skinned individuals. This skin color-based algorithmic bias was documented and widely publicized by the Gender Shade project that audited commercial facial recognition algorithms of major AI companies, including IBM and Microsoft (Buolamwini and Gebru, 2018). In the medical domain, scholars have also cautioned about the lack of transparency and patient diversity in the training datasets of AI algorithms for diagnosing skin cancer and diseases (Daneshjou et al., 2021). However, at the same time, facial recognition technology has been increasingly adopted by the beauty industry as skin analysis and diagnostic tools to recommend skincare products, such as L'Oréal's Skin Genius. To illuminate the effectiveness of XAI for facilitating bias detection, this study focuses on colorism, a system that favors those that possess lighter complexions, which is a prevalent bias in both the beauty industry and in facial recognition algorithms (Childs, 2022).

1.2. EXplainable AI

XAI is one of the major research fields for propelling the development of responsible AI (McDermid et al., 2021). As recent AI techniques such as deep learning are becoming more and more powerful, these models have also become too complex even for their developers to understand. XAI researchers aim to create technical solutions to make the AI model and its decisions understandable. It is important to note that XAI research focuses on explanations of AI decisions that are automatically generated by algorithms to describe the model's internal structure or output, not explanations retrofitted and manually crafted as an add-on for AI models. Via comprehensible explanations, accessible interfaces, and user-centered interactions, the ultimate goal of XAI is to help users understand why or why not an outcome has been derived from the system, when the system would succeed or fail, and when to trust the system or know it has erred (Gunning and Aha, 2019).

XAI researchers have proposed and tested various techniques (e.g., visualization of an AI model's behavior) and types of explanations, such as the "feature relevance" explanation that highlights what features in the input data lead to a particular AI output (for comprehensive reviews on XAI approaches, see Arrieta et al., 2020; Mohseni et al., 2021). Many people are familiar with XAI applications that recommend media content, such as news, and explain why this recommendation was offered to that particular person (Shin, 2021b). This study focuses on the popular "explanation by example" approach in XAI that is one of the longest-standing XAI techniques (Kenny et al., 2021) and also relatively easy to implement (Mohseni et al., 2021). More importantly, similar to the long-established use of case-based explanations in fields such as law and medicine, it has been consistently suggested by researchers that explanations-by-example are more intuitive and easier to understand than other explanation techniques (Montenegro et al., 2022). The growing user evaluation research in XAI has also generally supported the utility of this approach in improving user understanding of AI systems (Kenny et al., 2021).

This approach presents samples from the model's training dataset that are similar to the user's input in the model's representation space. This explanation type can feel more intuitive and easy to understand, as it is compatible with how humans often behave when attempting to explain a given process (Arrieta et al., 2020). We argue that this explanation type may be crucial to revealing the problem of a lack of diversity in the training datasets of AI systems, which is the most frequently discussed cause of algorithmic bias. As users can evaluate whether the examples provided are similar to or compatible with their own situations, the "explanation by example" approach may help users become aware of diversity and inclusivity issues in AI systems.

Critically, there exists scarce empirical data on the potential of XAI for helping users identify algorithmic bias. A notable exception is Melsión et al.'s (2021) experiment that used XAI visual explanation to help children understand gender bias in image classification algorithms. Their XAI visualization highlighted the parts of a given image that were important for the algorithm in determining the gender of the person in the image. The XAI labeled the person in an image as a woman when the image contained stereotypically "feminine" settings and objects, such as kitchen and cookware, and vice versa for a man if objects conventionally associated with masculinity were present. The analysis on the data collected from 84 preadolescents confirmed that the visual explanation helped these users understand and recognize algorithmic bias. Despite XAI's promising potential, however, little research attention has been directed toward understanding the mechanisms through which XAI may improve bias recognition and user trust.

Moreover, different stakeholders in AI systems have different XAI needs. Davis et al. (2020) distinguish two types of general users: *end users* versus *imposed users*. Unlike *end users*, who actively and voluntarily use AI as a tool to complete a task (e.g., using AI-powered design software to edit images), *imposed users* “are affected by the model’s decisions, outcomes, or recommendations” (p.3) without their volition (e.g., job applicants who have to pass AI resume screening). Many XAI studies that involved participants as non-expert *end users* often relied on personally irrelevant tasks. For example, in Ford and Keane’s (2022) study examining users’ domain expertise, participants were tasked with “debugging” an image classification program by judging the correctness of classifications with and without explanations and to report their ratings about the explanation helpfulness and overall satisfaction in the classification system. In another example, participants were asked to guess the age of the person in a picture with AI assistance (i.e., AI suggested the age but the participant made the final guess) (Chu et al., 2020). In other words, the AI decisions, as well as associated explanations, had no real-world impacts on the participants. In this way, the study results are not applicable to imposed users whose lives may be impacted by the AI output they received.

In stark contrast to the research attention on end users, *imposed users* are rarely studied, despite their vulnerability. With the rapid penetration of AI across industries and domains, more people will become imposed users, which could further disempower minority group members who are more likely to be negatively impacted by algorithmic bias (Ledford, 2019). Therefore, this study focuses on *imposed users* in an experimental setting that requires them to use AI’s visual analyses of their own faces via webcams in order to receive skincare recommendations. Recommendation systems are chosen because they represent one of the most common AI applications and one of the earliest and major domains of XAI adoption (Rong et al., 2023). Additionally, visual analysis using deep learning such as image and facial recognition is widely studied in XAI, while facial recognition is one of the most prominent examples for discussing AI bias and disparity.

End users often are domain experts who require and depend on information about the probability or confidence of AI decisions. In contrast, imposed users seek to assess whether the logic or rules applied to themselves were similarly applied to other groups to determine whether they may have been mistreated by the AI system (Kong et al., 2024). Recognizing the imposed users’ strong need to ascertain fairness and inclusiveness when interacting with AI systems, this study evaluates whether and how XAI influences imposed users’ fairness and inclusiveness judgments and the associated outcome of trust and bias detection.

2. Theoretical framework

2.1. XAI and perceived fairness

This study adopts the fairness heuristic theory (FHT) which has been recognized as an important theory and empirical framework for understanding the psychology of justice (Van den Bos and Lind, 2004). FTH has been applied in various settings including the workplace (Jordan et al., 2022) and the consumer market (Wang et al., 2021) to understand how people form and use fairness judgments as a proxy of interpersonal trust to guide decisions and behaviors (Lind, 2001). According to FHT, people undergo a multiphase process when forming and using fairness judgments. When dealing with new or uncertain situations (e.g., joining a new organization or interacting with an unfamiliar AI system), people rely on whatever justice-related information is at hand to establish an overall fairness assessment of the situation or entity, known as the “judgment phase.” People rely on their overall fairness judgment as a heuristic, defined as an efficient cognitive process (Gigerenzer and Gaissmaier, 2011) that functions as a mental shortcut to reduce complex cognitive tasks to simple mental operations (Shin et al., 2022a). In this “use phase,” fairness heuristics are employed to guide evaluations, attitudes, and actions in that situation, until an extraordinary event prompts people to reassess their fairness heuristics (Proudfoot and Lind, 2015).

Additionally, in the “judgment phase,” people often draw inferences about overall fairness based on different justice components, including distributive justice (fairness of outcomes), procedural justice (fairness of decision making procedures), and interactional justice, which itself consists of informational justice (timeliness, accuracy, and adequacy of explanations for decision processes and outcomes), and interpersonal justice (dignity and respect shown by authorities) (Beugre and Baron, 2001; Jones and Martens, 2009). In the context of XAI, explanations of AI decisions pertain to informational justice regarding whether accurate and clear explanations for a decision are provided in a timely manner (Cheng et al., 2022). Notably, Acikgoz et al. (2020) employed FHT to study how applicants who are in a vulnerable position of limited organizational knowledge may rely on quickly formed fairness heuristics as a basis for their attitudes toward AI-based interviewing. They found that organizations’ inability to provide explanations of how AI evaluates applicants’ performance during interviews leads to low perceived informational justice. Similarly, in the setting of algorithmic skin analysis, due to most people’s limited knowledge of such technology, users are likely to be in a vulnerable position where they feel uncertain about whether they can trust the AI analysis, and therefore, they are likely to rely on available fairness heuristics to interpret the outcome. Notably, Shin’s studies on news algorithms found that explanations about why specific news stories were recommended improves perceived fairness of the news analytics systems (Shin, 2021b), platforms (Shin et al., 2022b) and chatbots (Shin, 2022). We thus argue that automated explanations provided by the AI system address informational justice and constitute salient and readily available fairness-related information, enabling users to form an overall fairness assessment during the judgment phase. Perceived overall fairness of the AI system then functions as a heuristic to guide users’ responses to the system, including trust and bias recognition.

The reliance on fairness heuristics becomes even more pronounced when individuals find themselves in situations marked by high uncertainty (Al-Gasawneh et al., 2022). For instance, research suggests that when people have limited knowledge about a new technology they perceive as risky, they tend to rely on perceived fairness of the associated authorities when forming attitudes toward the technology (Song et al., 2021). As people are generally concerned about the risks and black-box nature of AI (Araujo et al., 2020; Brauner et al., 2023), fairness perceptions about AI systems are found to significantly impact user perception, attitude, and trust (Shin et al., 2022a; Shin, 2022; 2023a).

The idea that similar individuals should be treated similarly has been recognized as a common and intuitive principle of fairness (Dwork et al., 2012; Friedler et al., 2021). For instance, "treating similarly risky people similarly" has been recognized as a key aspect of algorithmic fairness (Zhou et al., 2020, p. 377). Kong et al. (2024) further suggest that imposed users are primarily concerned with whether they are being treated similarly to other similar users to form their fairness judgment.

When responding to explanations by example provided by AI systems, users are likely to evaluate whether the provided explanations accurately and adequately explicate the decision process and outcome (i.e., informational justice) to form their fairness perceptions. Therefore, in this study, users may establish an overall fairness judgment by evaluating whether the supposedly similar examples provided by the AI actually resemble themselves. To be specific, the depicted congruence between the user and the explanatory examples becomes a positive cue of fairness. In turn, the perceived congruence can positively affect users' perception of the AI system as being fair. In contrast, if the explanatory examples are noticeably incongruent—such as showing light-skinned examples to a dark-skinned user—this discrepancy violates information justice and may negatively impact the perceived overall fairness.

2.2. *Explanations by example and perceived inclusiveness*

Fairness is closely related to issues of inclusion and belonging (Lind, 2001). However, in sharp contrast to the growing research attention to fairness in AI systems, inclusiveness has not received adequate theorization and empirical examination. Inclusiveness in AI involves sundry considerations, from algorithms designed by diverse teams to ensure multiple perspectives, to representative training data ensuring that the algorithms are inclusive to different user needs (Avellan et al., 2020). As this study focuses on the widely observed problem of non-inclusive datasets causing algorithmic biases, we evaluate the significance of users' perceived inclusiveness of AI databases as another important psychological factor underlying the effects of XAI. FHT posits that people face two major threats in social relations–exclusion and exploitation (Lind, 2001). As information that helps individuals identify potential threats of exclusion can be a crucial decision heuristic, we contend that when users evaluate AI's explanations based on examples of similar users, they form judgments not only about the AI system's fairness but also its inclusiveness.

Perceived inclusiveness refers to feelings of being included and accepted in a given social situation (Chen and Tang, 2018). Originally developed in fields of education and social work to study diversity effects, perceived inclusiveness has been increasingly studied in management and organizational research (e.g., Shore and Chung, 2022). For instance, research based on social identity theory suggests that as employees identify with other members in the workplace, the perception of being included and appreciated by the identified social group positively impacts outcomes such as organizational citizenship behavior (Rice et al., 2021). From the theoretical perspective of social exchange, because employees value their involvement in the organization, perceived inclusion reinforces the quality of social exchange between the employee and organization and further induces compliance behavior (Chen and Tang, 2018).

Specific to the study focus on users' evaluation of AI explanations based on similar examples, we refer to social categorization theory to conceptualize how the explanation by example approach may influence inclusiveness perceptions toward the AI system. Social categorization theory suggests that people intuitively categorize themselves and others based on perceived similarity (i.e., ingroup members) and dissimilarity (i.e., out-group members) of observable physical characteristics, such as race, skin tone, age, and gender (Turner, 1987). Skin tone constitutes an explicitly observable and essential basis for in- and out-group categorization (Uzogara et al., 2014). Additionally, as skin tone is a decisive factor for different skin issues and needs, participants in our study's setting of AI skin analysis are predisposed to be highly cognizant of the similarity of skin tone between the provided examples and themselves. As a result, when users notice that the supposedly similar others in the AI system's explanation are in fact quite different from themselves (i.e., dramatically different skin tone), they are likely to infer that there are no users like themselves (i.e., ingroups) included in the AI system's database, resulting in feelings of being excluded and questioning the inclusiveness of the AI system. Research on workplace diversity also shows that perceived similarity or dissimilarity with others in social environments can affect individuals' perceived inclusiveness. For instance, Bae et al. (2017) found that employees' demographic dissimilarity (i.e., gender and race) with their supervisor and coworkers is negatively associated with perceived organizational inclusion. Similarly, Şahin et al. (2019) observed that felt inclusion was lower among individuals who perceived themselves as dissimilar to others at work at a deep level, compared to those who perceived themselves as similar. Thus, the following hypothesis is posited:

> **H1**: Participants in the skin tone congruency condition (vs. incongruency condition) will (a) perceive the AI system as fairer and (b) consider the system as more inclusive.

2.3. *The mediating roles of perceived fairness and inclusiveness*

As aforementioned, overall fairness assessments serve as cognitive shortcuts that influence subsequent various responses, including user engagement (Roy et al., 2028) and satisfaction (Wang et al., 2021) in the "use phase" (Lind, 2001). Conversely, if the initial perception is one of unfairness, the use phase is marred by negative reactions.

Recent research on AI fairness has also documented a strong correlation between perceived fairness and user trust (Shin, 2021a, 2023a; Shin et al., 2022b). Therefore, perceived fairness is expected to mediate the effects of skin tone (in)congruence on trust. That is, when users find the explanatory examples to be congruent to their self-images (e.g., congruent skin tone) and thus deem the AI system as fair, their trust in the AI system is likely to increase. Conversely, the mismatch in skin tone in the system's explanations is likely to be interpreted as a sign of unfairness, thereby decreasing their trust.

Moreover, given that fairness is often linked to perceptions of justice and absence of bias, experiencing feelings of unfairness could make individuals more alert to potential biases. Research on organizational justice suggests that the relevant concepts of procedural and distributive justice mediate the effects of promoting a minority (e.g., promoting a woman into management) on the associated discrimination perception (e.g., gender discrimination) (Russen et al., 2021). That is, positive justice perceptions decrease people's belief in discrimination. Along this line of reasoning, fairness perceptions are expected to influence users' perceived bias in AI systems. Hence, we hypothesize that perceived fairness of the AI system will mediate the relationship between skin tone (in)congruence and awareness of algorithmic bias.

> **H2**: The impact of skin tone (in)congruency on (a) trust in the AI system and (b) awareness of algorithmic bias will be mediated by perceived fairness of the system.

Inclusiveness has been increasingly emphasized by scholars as a key component of responsible AI (Zhou et al., 2020; Schelenz et al., 2021). We argue that, in addition to fairness, perceived inclusiveness constitutes another critical decision heuristic for guiding user response. In the workplace setting, research indicates that perceptions of inclusion can enhance the trusting climate among employees (Downey et al., 2015) and fosters employees' commitment to their organizations (Chen and Tang, 2018). The related literature on social exclusion further suggests that the greater experience with social exclusion is associated with lower general trust toward others. As such, we expect that perceived inclusiveness will mediate the effects of skin tone (in)congruence on trust. Additionally, research suggests that social exclusion and perceived discrimination are positively associated (Saasa, 2019). Therefore, experiencing feelings of exclusion could make individuals more sensitive to possible biases in AI systems. Accordingly, we posit that perceived inclusiveness of the AI system will mediate the impact of skin tone (in)congruence on awareness of AI biases.

> **H3**: The impact of skin tone (in)congruency on (a) trust in the AI system and (b) awareness of algorithmic bias will be mediated by perceived inclusiveness of the system.

2.4. The moderating role of experiences with colorism

Focusing on algorithmic bias in relation to skin tone, we evaluated the moderating role of a user's lived experience with colorism in the relationships between skin tone (in)congruency and perceived fairness and inclusiveness of the AI system, respectively. In the organizational justice literature where FHT has been widely studied, important individual differences, such as trust propensity and risk aversion (Colquitt et al., 2006), have been identified as moderators of injustice effects. Individuals high in justice sensitivity tend to perceive and recall unjust events more frequently (Schmitt and Dörfel, 1999). More importantly, they tend to develop highly accessible injustice concepts that shape their attention, interpretation, and memory for justice-related information (Baumert et al., 2011). However, justice sensitivity is conceptualized as a personality trait, and the literature has not sufficiently explored how individuals' lived experiences with injustice and marginalization might influence their reaction to justice and fairness-related information. A notable exception is Dierckx et al.'s (2023) study on ethnic-cultural decision making. They observed that minority group members respond to (un)fair treatment of a different minority group on the basis of shared disadvantaged group membership. Recognizing the disproportionate adverse impacts of algorithmic bias on minority groups, this study expands FHT by assessing how people's varied experiences with discrimination as an individual difference factor may influence how they respond to XAI to form fairness perceptions.

Regarding inclusiveness perceptions, based on social identity theory, research suggests that social minorities tend to feel a lower level of inclusiveness as they have been historically marginalized in social interactions (Cho and Mor Barak, 2008). Studies on workplace inclusion further suggest that employees representing a numerical minority are constantly gathering and assessing information about who belongs and who does not (Şahin et al., 2019). As minority employees are acutely sensitive to social acceptance cues in the workplace, cues such as all-white conference speakers or displays of all male leaders, can lower feelings of inclusiveness among minority employees who are dissimilar to those cues (Latu et al., 2013; Murphy et al., 2018). Accordingly, it is expected that users who have experienced discrimination based on skin color (i.e., colorism) will react more strongly to explanatory examples of other individuals of incongruent skin tone when forming fairness and inclusiveness perceptions.

Specific to the study outcome of raising bias awareness, research shows that experiencing discrimination can trigger individuals' recognition of inequality and injustice in social structures (Curtin et al., 2015). The impacted individuals tend to be mindful of their marginalized status and become more vigilant about bias and prejudice against people like them (Jun et al., 2021). For instance, Van Prooijen et al. (2004) observed that people's experiences with being socially excluded moderate their reactions to unrelated experiences of procedural justice. Additionally, Jun et al. (2021) found that Asian Americans' past experience with racism is positively associated with their problem recognition of the recent anti-Asian discrimination during the COVID-19 pandemic. It is important to note that, in recent years, Black social media influencers and consumers have called out colorism and the relative invisibility of darker-skinned models in the beauty industry (Childs, 2022). Therefore, this study explores whether users' past experiences with colorism may augment their perceptions of the AI system.

> **H4**: Past experience with colorism will moderate the effects of skin tone (in)congruency on perceived fairness of the AI system such that the effects will be more salient among participants with more experiences with colorism.
> **H5**: Past experience with colorism will moderate the effects of skin tone (in)congruency on perceived inclusiveness of the AI system such that the effects will be more salient among participants with more experiences with colorism.
> **H6**: Perceived fairness of the AI system will mediate the interaction effect between skin tone (in)congruency and experience with colorism on (a) trust in the AI system, and (b) awareness of algorithmic bias.

H7: Perceived inclusiveness of the AI system will mediate the interaction effect between skin tone (in)congruency and experience with colorism on (a) trust in the AI system, and (b) awareness of algorithmic bias.

3. Method

To validate the hypotheses, a one factor between-subject experiment was carried out among 241 U.S. female consumers via CloudResearch Connect and Amazon Mechanical Turk. Skin tone (in)congruency was the manipulated factor.

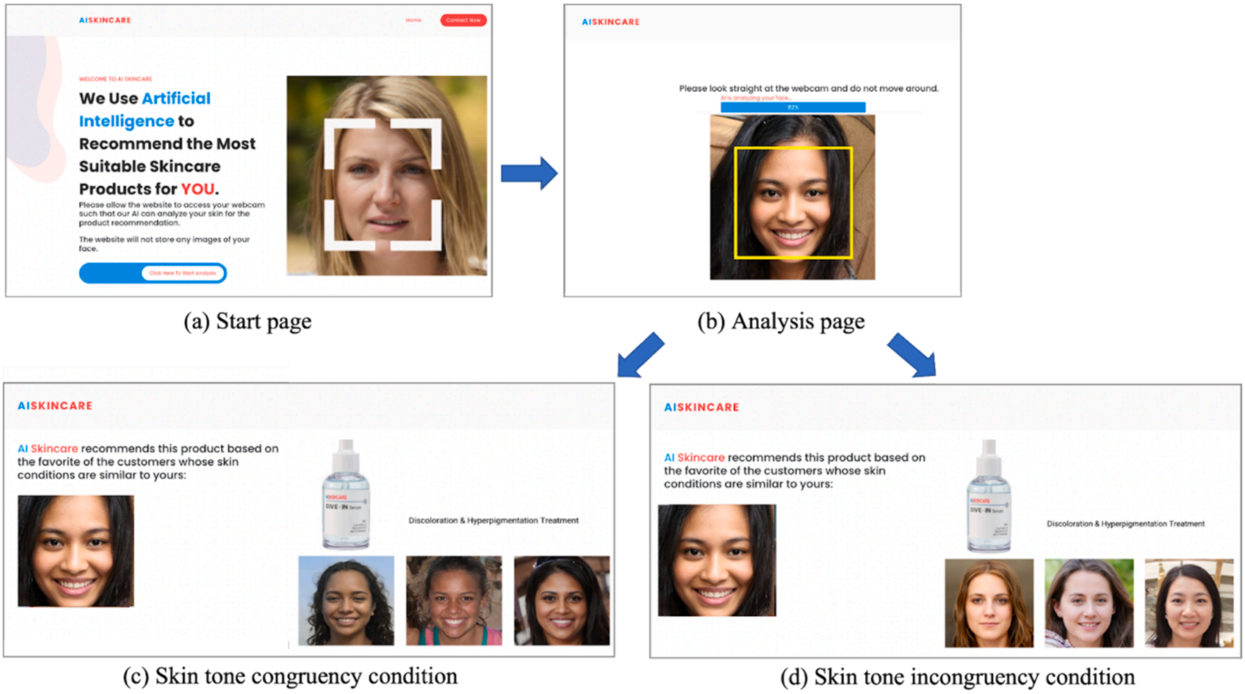
3.1. Stimuli development

In this study, the AI-powered skincare recommendation system (AI Skincare) will recommend a product to the user (i.e., the participant) that is favored by three other users whom the AI system identified as having similar skin analysis results to the participant. On the product recommendation page, alongside the recommended product, the participant can see images of the three AI-identified similar users.

A fictitious AI skincare website was created as the stimulus website (see Fig. 1). Upon arrival on the homepage, participants were greeted with a brief introduction: “Welcome to AI Skincare! We use artificial intelligence to recommend the most suitable skincare products for you.” (Fig. 1a). Simultaneously, on this page, participants were instructed to activate their webcams to initiate the AI skincare analysis. Once they clicked “Start Analysis,” participants were directed to a page (Fig. 1b) that allowed them to interact with an AI system designed to scan and analyze their faces using their webcams. On the results page, participants received an AI-recommended product, based on their skin analysis, along with an explanation for the AI recommendation (Fig. 1c and d). The AI system explained that the product was recommended as it was popular among other users with similar skin analysis results. Participants could also see their own face in a photo captured during the analysis process and the images of three exemplar users. The exemplar user images, though highly realistic, were fictitious faces crafted using generative adversarial networks (Karras et al., 2020).

The congruency between the AI-provided photos and the participant was manipulated according to the participant’s self-reported skin tone (see the study procedure described later). Participants in the congruent condition received an AI-recommended skincare product based on the preferences of other users sharing the same skin tone. Specifically, participants with darker skin tones were shown photos of other darker-skinned users (see Fig. 1c), while those with lighter skin tones saw photos of other lighter-skinned users. Conversely, participants in the incongruent condition were shown photos of users with contrasting skin tones. For instance, participants with darker skin tones saw photos of lighter-skinned users who were nonetheless categorized as similar by the AI system (Fig. 1d).

We chose face serum as the focal product, given its prevalence as a common skin care item. As skin issues are often correlated with age, we created three versions of the serum product, each targeting the main skin problems of female consumers within different age



We Use Artificial Intelligence to Recommend the Most Suitable Skincare Products for YOU.

Please allow the website to access your webcam such that our AI can analyze your skin for the product recommendation.

The website will not store any images of your face.

(a) Start page

Please look straight at the webcam and do not move around.

AI is analyzing your face...

(b) Analysis page

AI Skincare recommends this product based on the favorite of the customers whose skin conditions are similar to yours:

(c) Skin tone congruency condition

AI Skincare recommends this product based on the favorite of the customers whose skin conditions are similar to yours:

(d) Skin tone incongruency condition

Fig. 1. The AI Skincare website designed for the experiment stimuli.

groups. The serum was designed to provide discoloration and hyperpigmentation treatment for consumers aged 18–31, brightening and wrinkle care for consumers aged 31–50, and advanced anti-aging care for consumers aged 51 and above. Participants in different age groups received different product recommendations from the AI system. In addition, to ensure the age alignment between the participant and the fictitious faces of the simulated exemplars, we created a machine learning model using transfer learning on Resnet (He et al., 2016) with the IMDB-WIKI dataset (Rothe et al., 2015) to determine the age of the fictitious face. In this way, participants will see faces of users in their age group, but in one of the two conditions regarding skin tone congruence.

3.2. Participants

Given that the context of our study was focused on an AI skin analysis for skincare recommendations, we solely recruited female participants as women are more than two times as likely to have a skincare routine than men (62 % vs. 29 %) (Schriber, 2023). A total of 241 valid responses were collected from female consumers in the U.S., aged between 18–79, after excluding participants who submitted incorrect codes, shared the same IP address, failed the attention check questions, or did not activate their webcams. Of these, 80 participants (33.2 %) were in the 18–30 age range, 105 (43.6 %) fell into the 31–50 range, and 56 (23.2 %) were aged 51 or above. Among these participants, 105 (43.6 %) self-identified as light-skinned, while 136 (56.4 %) considered themselves dark-skinned. Many were African Americans (46.5 %), followed by White (37.3 %), Asian Americans (6.2 %), Hispanic Americans (3.7 %), mixed race (4.6 %), and Other (1.6 %). In terms of educational background, the majority held a 4-year college degree (42.7 %). This was followed by those with some college education (19.5 %), associate degrees (12.9 %), Master's degrees (12.0 %), high school diplomas (9.5 %), and other forms of education (3.3 %). Each was paid $5.00 for their participation.

3.3. Procedure

Participants across the three age groups (18–30, 31–50, 51 and above) were initially asked to indicate their skin tone and report their familiarity with AI, along with their general trust in AI. Within their respective age groups, each participant was then randomly assigned to either the skin tone congruent or incongruent condition. After that, participants' perceptions of the AI Skincare website, their evaluation of the AI system's explanations, and their personal experiences with colorism were assessed through a brief survey. At the end of the experiment, all the participants were debriefed about the study purpose and manipulation, including the deception about the AI skin analysis, product recommendation, the computer-generated images of similar users, and why the deception was necessary. The study protocol was approved by the institutional review board at the authors' institution.

3.4. Measures

Wherever possible, we sourced measurement items for the constructs from previously validated scales. All variables were assessed on a 7-point scale. To verify the success of our experimental manipulation of the skin tone congruency between the participants and the simulated exemplars, we asked participants to rate their agreement with the skin tone congruence with the similar users identified by the AI system.

Measurement items of trust in the AI system (a = .96) were adopted from Shin (2021a). We employed seven items, including statements like "I trust the recommendations by AI Skincare" and "Recommended items by AI Skincare are trustworthy." To measure awareness of algorithmic bias (a = .88), we asked participants to report their awareness level of the various ways in which an AI system can exhibit bias. We created five items to measure participants' level of awareness algorithmic bias in general, based on the different types of algorithmic bias conceptualized by Danks and London (2017), including "Algorithmic focus bias: the system is built by focusing on either irrelevant or sensitive information such as race from the data" and "Training data bias: the data used to build the system is biased. For example, the data does not have the same representation for each group of people."

Perceived fairness (a = .94) was measured by three items adopted from Shin (2021a). Example items include "AI Skincare shows no favoritism and does not discriminate against people"' and "I believe the design of AI Skincare follows due process of impartiality with no prejudice." Perceived inclusiveness (a = .85) was assessed by asking participants to indicate how inclusive they perceived the AI Skincare database to be. Two items were created: "To what extent do you think the database of AI Skincare is inclusive?" and "To what extent do you think the database of AI Skincare includes people like you?". Direct experience with colorism (a = .87) was measured with four items adapted from Sanchez and Brock (1996) and Shaffer et al. [85]. Example items are "I do not get enough recognition because of my skin color" and "I have been unfairly treated because of my skin color.".

Three control variables, AI familiarity, general trust in AI, and perceived accuracy of the AI skin analysis result, were also measured. AI familiarity (a = .91) was assessed via four items adopted from Chi et al. (2021), such as "I know a lot about AI" and "I am familiar with AI." General trust in AI (a = .93) was measured by four items from Liu (2021) and included items such as "If a decision is made by AI, it must be precise" and "If a judgment is made by AI, it must be objective." Lastly, perceived AI analysis accuracy (a = .94) was created for this study and measured using one item of "How accurately does the result reflect your skin problem?".

4. Results

4.1. Manipulation check

To validate our manipulation of the skin tone congruency between the participants and the simulated exemplars, an independent-samples *t*-test was conducted. The analysis showed a significant effect of the skin tone congruency manipulation ($t(239) = 9.54$, $p < .001$). Participants in the congruent condition were more likely to agree that the AI-identified similar users had similar skin tones to them ($M = 5.08$, $SD = 1.50$) than those in the incongruent condition ($M = 3.13$, $SD = 1.68$), suggesting the success of our manipulation of skin tone congruency.

4.2. Hypotheses testing

To test H1, we conducted the analysis of covariance (ANCOVA), controlling for perceived AI analysis accuracy, AI familiarity, and general trust in AI. The analyses revealed significant main effects of skin tone (in)congruency on perceived fairness ($F(1, 236) = 25.52$, $p < .001$) and inclusiveness of the AI system ($F(1, 236) = 52.09$, $p < .001$). Specifically, participants in the congruent condition perceived higher fairness of the AI system ($M = 4.97$, $SE = .13$) and inclusiveness ($M = 5.32$, $SE = .13$) than those in the incongruent condition (perceived fairness: $M = 4.03$, $SE = .13$; perceived inclusiveness: $M = 3.96$, $SE = .13$). Thus, both H1a and H1b were supported.

To test H2 and H3, we used the simple mediation model (i.e., Model 4) of PROCESS in SPSS with bootstrapping of 5,000 samples (Hayes, 2017). Skin tone (in)congruency was the independent variable. Perceived fairness and inclusiveness of the AI system functioned as the mediator, respectively. Trust in the AI system and awareness of algorithmic bias served separately as the dependent variables. Perceived AI analysis accuracy, AI familiarity, and general trust in AI were included in the model as covariates. As shown in Table 1, perceived fairness of the AI system significantly mediated the influence of skin tone (in)congruency on trust and awareness of algorithmic bias. That is, when participants saw that the similar users identified by the AI skincare system did not have similar skin tones to them, they perceived the system as less fair, which in turn led to lower trust and higher awareness of algorithmic bias. Therefore, H2 was supported.

Moreover, as displayed in Table 2, perceived inclusiveness of the AI system significantly mediated the influence of skin tone (in) congruency on the two dependent variables. That is, when participants saw that the supposedly similar users did not match their skin tone, they perceived the system to be less inclusive, which again led to lower trust and higher awareness of algorithmic bias. Hence, H3 was also supported.

To test H4 and H5, PROCESS Model 1 was used. Skin tone (in)congruency was the independent variable; experience with colorism was the moderator; perceived fairness and inclusiveness of the AI system served as the dependent variables, respectively; perceived AI analysis accuracy, AI familiarity, and general trust in AI were the covariates. To prepare for the moderation analysis, experience with colorism was standardized. First, we found a significant two-way interaction effect between skin tone (in)congruency and experience with colorism on perceived fairness of the AI system ($B = .35$, $SE = .18$, $p = .04$). That is, the impact of skin tone (in)congruency on perceived fairness was significantly moderated by experience with colorism. As shown in Table 3, the effect of skin tone (in)congruency on perceived fairness was strengthened when participants had greater experience with colorism, supporting H4. Results indicated that the effect of skin tone (in)congruence on perceived inclusiveness was also significantly moderated by experience with colorism (two-way interaction effect: $B = .54$, $SE = .18$, $p = .002$). As seen in Table 4, the effect of skin tone (in)congruency on perceived inclusiveness of the AI system became stronger when participants had more experience with colorism, supporting H5.

In H6 and H7, we hypothesized that by influencing perceived fairness and inclusiveness, skin tone (in)congruency, along with experience with colorism, would ultimately affect trust in the AI system and awareness of algorithmic bias. We used the moderated mediation model (i.e., Model 7) of PROCESS, in which the above data analysis set-ups were integrated to test these two hypotheses.

Results confirmed the hypothesized moderated mediation effect via perceived fairness of the AI system (see Table 5). That is, skin tone (in)congruence and experience with colorism influenced trust in the AI system and awareness of algorithmic bias as a result of their interaction effects on perceived fairness. Moreover, the indirect effects of skin tone (in)congruency on the two dependent variables via perceived fairness was greater when participants had experienced greater colorism. H6 was thus supported. The moderated mediation via perceived inclusiveness of the AI system was also found. The indirect effects of skin tone (in)congruency on the two dependent variables via perceived inclusiveness was more prominent when participants had experienced greater colorism (see Table 6). Hence, H7 was supported.

5. Discussion

The widespread adoption of AI, driven by industry-wide mass automation, has raised critical concerns about algorithmic bias and its far-reaching adverse impacts on societal well-being, including the accelerated exacerbation of social inequities. Therefore, raising awareness on algorithmic bias and helping users recognize such biases in their everyday encounter with AI systems is the critical first step in mitigating algorithmic bias. While XAI has been suggested by scholars as a promising means to help people identify algorithmic

Table 1
Effects of skin tone (in)congruency as mediated by perceived fairness of the AI system.

	Direct effect of skin tone congruency	Effect of perceived fairness of the AI system	Indirect effect of skin tone congruency via perceived fairness
Trust in the AI system	$B = .29$, $SE = .15$, $p = .05$	$B = .26$, $SE = .05$, $p < .001$	$B = .25$, Boot $SE = .07$, 95 %CI [.12,.39]
Awareness of algorithmic bias	$B = -.20$, $SE = .20$, $p = .32$	$B = -.26$, $SE = .07$, $p < .001$	$B = -.24$, Boot $SE = .09$, 95 %CI [−.45, −.09]

Table 2
Effects of skin tone (in)congruency as mediated by perceived inclusiveness of the AI system.

	Direct effect of skin tone congruency	Effect of perceived inclusiveness of the AI system	Indirect effect of skin tone congruency via perceived inclusiveness
Trust in the AI system	$B = .26$, $SE = .16$, $p = .10$	$B = .20$, $SE = .05$, $p < .001$	$B = .27$, Boot $SE = .08$, 95 %CI [.14,.43]
Awareness of algorithmic bias	$B = -.23$, $SE = .22$, $p = .29$	$B = -.16$, $SE = .07$, $p = .02$	$B = -.22$, Boot $SE = .11$, 95 %CI [−.44, −.03]

Table 3
Effects of skin tone (in)congruency on perceived fairness of the AI system at different levels of experience with colorism.

Level of experience with colorism	Effect	*SE*	*p*
Low	.52	.24	.03
Medium	.87	.18	<.001
High	1.23	.26	<.001

Table 4
Effects of skin tone (in)congruency on perceived inclusiveness of the AI system at different levels of experience with colorism.

Level of experience with colorism	Effect	*SE*	*p*
Low	.76	.24	.002
Medium	1.30	.18	<.001
High	1.85	.26	<.001

Table 5
Indirect effects of skin tone (in)congruence via perceived fairness of the AI system on dependent variables at different levels of experience with colorism.

Dependent variable	Level of experience with colorism	Effect	Boot *SE*	Boot LLCI	Boot ULCI
Trust in the AI system					
	Low	.13	.07	.004	.29
	Medium	.23	.06	.11	.36
	High	.32	.09	.17	.51
	Moderated mediation	Index	Boot *SE*	Boot LLCI	Boot ULCI
		.09	.05	.001	.20
Awareness of algorithmic bias					
	Low	−.13	.08	−.31	−.002
	Medium	−.23	.08	−.40	−.09
	High	−.32	.11	−.55	−.12
	Moderated mediation	Index	Boot *SE*	Boot LLCI	Boot ULCI
		−.09	.05	−.21	−.001

Table 6
Indirect effects of skin tone (in)congruence via perceived inclusiveness of the AI system on dependent variables at different levels of experience with colorism.

Dependent variable	Level of experience with colorism	Effect	Boot *SE*	Boot LLCI	Boot ULCI
Trust in the AI system					
	Low	.15	.06	.05	.30
	Medium	.26	.07	.13	.42
	High	.37	.10	.18	.60
	Moderated mediation	Index	Boot *SE*	Boot LLCI	Boot ULCI
		.11	.05	.03	.21
Awareness of algorithmic bias					
	Low	−.12	.07	−.29	−.01
	Medium	−.21	.10	−.42	−.02
	High	−.29	.14	−.58	−.04
	Moderated mediation	Index	Boot *SE*	Boot LLCI	Boot ULCI
		−.09	.05	−.20	−.01

bias (Saeed and Omlin, 2023), this study provides some of the earliest empirical evidence on its effectiveness for this important but under-researched goal.

While the emerging XAI research stream in social sciences has focused on the broad value of AI explainability (e.g., Shin, 2021a), our research delves deeper by zeroing in on the widely adopted explanation-by-example approach and illuminates its importance in not only trust formation in users' interactions with AI systems but also in facilitating bias recognition. The results confirmed that this XAI approach enables users to assess whether the examples are similar and compatible to their inputs or circumstances. The alignment of the explanatory example with users' individual circumstances shape their perceptions of fairness and inclusiveness of the AI system. Conversely, when a disparity exists between the user and the explanatory examples, it serves as a red flag, signaling unfairness and exclusion, thereby prompting users not to put blind trust in the system and enabling users to recognize algorithmic bias stemming from non-inclusive datasets. Our results based on the explanation-by-example approach thus shed valuable insights on the pivotal question of how XAI can be leveraged to spotlight algorithmic bias.

5.1. *Theoretical implications*

By contextualizing the study within the FHT framework, our findings shed important theoretical insights on the psychological mechanism underlying the effects of XAI via perceived fairness and inclusiveness. The results illuminate the importance of fairness and inclusiveness perceptions for driving user response to AI systems. Specifically, considering the high level of uncertainty toward AI applications (Araujo et al., 2020; Brauner et al., 2023), explanations about AI recommendations provided salient and readily accessible information in relation to informational justice to address imposed users' concern about fairness of the AI system. Additionally, consistent with FHT and prior research on AI fairness (e.g., Shin 2020b; Shin and Park, 2019), perceived fairness of AI systems significantly impacted user trust. The results further highlight the crucial role of perceived fairness in facilitating bias recognition that has not been studied in the literature. That is, when an AI system is perceived as unfair, users are more likely to pay attention to potential biases demonstrated by the AI system. The results thus expand the explanatory power of FHT by demonstrating its utility in the context of user response to XAI and for the outcome of bias recognition. This finding also reaffirms scholars' recommendation to recognize fairness perceptions as a crucial element in understanding the broader implications of algorithmic fairness (Starke et al., 2022).

Expanding prior research on AI explainability by focusing on the explanation by example approach, the findings highlight the under-researched factor of perceived inclusiveness in the formation of trust and development of bias awareness. The results demonstrated that this XAI approach enables users to easily and quickly assess whether the AI system's dataset includes similar users like themselves to inform their evaluation of the inclusiveness of the system, which in turn affect their trust and bias awareness. Resonating with scholars who have advocated for the significance of inclusiveness in the development and employment of AI systems (De Cremer and De Schutter, 2021), this study calls for more empirical and interdisciplinary research on the characteristics of inclusive AI as well as the factors driving inclusiveness perceptions toward AI systems.

Furthermore, the unique focus of this study on imposed users sets it apart from prior XAI research that has predominantly centered on end users. It is important to note that algorithms with "humans-in-the-loop" have been proposed as technical solutions to mitigating algorithmic bias (Rodrigues, 2020). However, in such human-in-the-loop machine learning, controls are typically given to expert users (Mosqueira-Rey et al., 2023), not the imposed users. While imposed users represent a rapidly growing stakeholder group in the AI ecosystem, they remain under-studied. By the same token, social minorities who are the most vulnerable to the negative impacts of algorithmic bias have not received sufficient research attention in XAI research. Our study findings addressed this notable gap by underscoring the potential of XAI for informing and empowering imposed users, particularly those belonging to marginalized groups, for evaluating and scrutinizing the fairness and inclusiveness of AI systems that are now increasingly making decisions that impact their lives.

Our findings also expanded FHT by highlighting the importance of experience with discrimination as a critical individual difference in the formation and use of decision heuristics of fairness and inclusiveness. Prior research suggests that past experiences with discrimination can lead to an increased activation of discrimination-related thoughts when people encounter ambiguous situations, and these thoughts, in turn, can influence their interpretations of these circumstances (Van Prooijen et al., 2004). As fairness is closely related to issues of discrimination, inequality, and exclusion, our study provided much needed empirical evidence on how individuals' varied discrimination experiences may impact how sensitive they are toward information on fairness and inclusiveness. Specifically, participants with greater experiences with colorism, upon noticing the incongruent skin tone of the reportedly similar users, demonstrated stronger perceptions of unfairness and non-inclusiveness of the AI system, compared to those with little or no experience with colorism. Therefore, beyond the organizational justice setting studied in prior research, our findings confirmed the moderating effects of prior experience with discrimination on fairness and inclusiveness perceptions in the XAI context.

5.2. *Practical implications*

The findings of our study attest to the instrumental value of XAI in helping users understand the decision-making process of AI, thereby shaping their perceptions toward and trust in the AI system. These results can serve as a compelling impetus for legislators to craft policies that can facilitate and expedite the integration of XAI into the machine learning workflow. In particular, the results confirmed that the explanation by example approach effectively enabled participants to recognize algorithmic bias emanating from non-inclusive datasets and prevented blind trust in such biased AI systems. For high-stake AI applications, such as loan approval and resume screening, this XAI technique has the potential to assist various stakeholders in performing algorithmic auditing to mitigate and prevent the dire consequences of algorithmic bias. Additionally, the results revealed that when the explanatory examples are congruent or compatible to users' personal circumstances, XAI can improve perceived fairness and inclusiveness to increase user trust. As research has documented the significance of trust for enhancing important outcomes such as perceived usefulness and convenience (Shin, 2020b) and continuance intention of the AI system (Shin, 2020a), organizations may capitalize on this XAI approach to enhance user trust.

Additionally, advocacy and civil rights groups (e.g., The Center for Civil Rights and Technology by The Leadership Conference on Civil and Human Rights) may consider incorporating this XAI approach in their educational campaigns to boost AI literacy and enhance understanding about algorithmic biases, particularly among non-expert users. Relatedly, companies employing AI in their internal decision-making processes or embedding AI in their services and products should incorporate XAI to facilitate algorithmic auditing. While this study focused on imposed users, XAI may also be helpful for professional users to ensure responsible usage of this powerful technology, particularly for consequential decisions (e.g., recommend a medical treatment). We argue that incorporating XAI is essential for organizations to practice and demonstrate a commitment in ethical decision-making in the AI era. (Roovers, 2019).

5.3. Conclusion

Focusing on the explanation by example approach, this research outlines how XAI enabled imposed users, one of the most vulnerable stakeholder groups in the AI ecosystem, to recognize one of the most widely discussed algorithmic bias resulting from non-inclusive datasets. By contextualizing this XAI user study within theoretical frameworks of fairness heuristics theory and social categorization theory, the results provided much needed empirical insights on the psychological mechanism underlying the XAI effects via perceived fairness and inclusiveness. A congruence between the explanatory example and user expectations signals the fairness and inclusiveness of the AI system, while a discrepancy serves as a warning of potential unfairness and exclusion. In turn, user perceptions of fairness and inclusiveness are crucial mechanisms for building trust towards the AI system and enhancing awareness of algorithmic bias. Additionally, users' prior experience with discrimination emerged as a significant boundary condition. XAI elicited stronger reactions from users who experienced greater discrimination than those with minimal discrimination experience. These insights not only deepen our understanding of the social impacts of XAI in relation to algorithmic bias but also provide valuable guidance for companies using these technologies and policymakers regulating AI.

Researchers have speculated on future capacities of XAI for not only explaining AI decisions to individual users but also enabling users to interact with explanations to provide feedback and steer learning (Gunning et al., 2019). Human-centered XAI which addresses the importance of different stakeholder groups and their varied needs (Ehsan and Riedl, 2020) is imperative. This study thus joins the ongoing call for more interdisciplinary research (Johs et al., 2022) to provide human-centered guidance on developing and evaluating XAI approaches for mitigating algorithmic bias (Manresa-Yee et al., 2021).

5.4. Limitations and future research directions

Several limitations should be considered and addressed in future research. First, although this study provides valuable empirical evidence on the potential of XAI for impacting user perceptions of fairness and inclusiveness and raising awareness of algorithmic bias, awareness and perceptions are only the first step to mitigating algorithmic bias. Future studies should explore other important outcomes, such as attitude toward AI policies and regulations. Future research should also evaluate behavioral outcomes, such as sharing and publicizing one's own experiences with algorithmic biases for crowdsourced/collaborative algorithm auditing (Shen et al., 2021). Second, the reliance on a single experiment with a small, homogenous sample points to the need for further inquiry. To enhance the validity and robustness of the findings, additional experiments that employ different XAI approaches, distinct study contexts, and diverse samples, are needed. For instance, the current study only focused on the popular and easily implemented explanation-by-example XAI approach, while there exist various XAI types and approaches. Future research thus should explore different XAI approaches to compare their utility for different purposes, from bias detection to trust formation and development of AI literacy. Furthermore, this study explored XAI only in the context of the facial recognition algorithms and the commonly associated algorithmic bias of colorism. As Shen et al. (2021) noted that individuals may have varying levels of sensitivity to distinct types and formats of bias, more research attention is needed to explore the effects of XAI on detection for different types of biases across different AI applications. In a similar vein, the participant demographic in this study was limited to females. We acknowledge that the female-only sample overlooks the potential interest and utility of AI-powered skin analysis for male and transgender individuals, thereby introducing a sampling bias. Moving forward, it is necessary to involve more diverse study participants to validate the generalizability of our findings. Relatedly, while this study focused on imposed users whose lives and decisions are directly impacted by the AI system, the experiment setting of AI for skincare recommendation involves relatively low stakes. As fairness perceptions of algorithm decision-making are highly context-dependent (Starke et al., 2022), future studies should explore the effects of XAI across distinct contexts (e.g., high and low stakes, positive and negative outcomes) (Leichtmann et al., 2023), and examine whether users process the explanation through the central route or peripheral route according to dual-process theories such as Elaboration Likelihood Model and Heuristic-systematic model (e.g., Shin, 2020a). If the explanation is processed peripherally or mindlessly, XAI can lead to blind trust, thus demanding more critical attention. Finally, employing a multi-experiment design with repeated measures could further validate the study's conclusions by ruling out reverse causality in perception-related constructs. With additional experiments, possibility of reverse causation can be removed if consistent empirical evidence appears. Researchers can also explore possible alternative explanations and identify additional confounding variables to further improve study validity.

CRediT authorship contribution statement

Ching-Hua Chuan: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Ruoyu Sun:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Shiyun Tian:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization. **Wan-Hsiu Sunny Tsai:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Acikgoz, Y., Davison, K.H., Compagnone, M., Laske, M., 2020. Justice perceptions of artificial intelligence in selection. Int. J. Sel. Assess. 28 (4), 399–416.

Al-Gasawneh, J., Alfityani, A., Al-Okdeh, S., Almasri, B., Mansur, H., Nusairat, N., Siam, Y., 2022. Avoiding uncertainty by measuring the impact of perceived risk on the intention to use financial artificial intelligence services. Uncertain Supply Chain Manage. 10 (4), 1427–1436.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. 2016. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed October 24, 2023).

Araujo, T., Helberger, N., Kruikemeier, S., De Vreese, C.H., 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. 35, 611–623.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Avellan, T., Sharma, S., Turunen, M., 2020. AI for All: Defining the what, why, and how of inclusive AI. In: Proceedings of the 23rd International Conference on Academic Mindtrek, pp. 142–144.

Bae, K.B., Sabharwal, M., Smith, A.E., Berman, E., 2017. Does demographic dissimilarity matter for perceived inclusion? Evidence from public sector employees. Rev. Public Person. Admin. 37 (1), 4–22.

Baumert, A., Gollwitzer, M., Staubach, M., Schmitt, M., 2011. Justice sensitivity and the processing of justice–related information. Eur. J. Pers. 25 (5), 386–397.

Beugre, C.D., Baron, R.A., 2001. Perceptions of systemic justice: The effects of distributive, procedural, and interactional justice. J. Appl. Soc. Psychol. 31 (2), 324–339.

Brauner, P., Hick, A., Philipsen, R., Ziefle, M., 2023. What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. Front. Comput. Sci. 5, 1113903.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

Chen, C., Tang, N., 2018. Does perceived inclusion matter in the workplace? J. Manag. Psychol. 33 (1), 43–57.

Cheng, J., Usman, M., Bai, H., He, Y., 2022. Can authentic leaders reduce the spread of negative workplace gossip? The roles of subordinates' perceived procedural justice and interactional justice. J. Manag. Organ. 28 (1), 9–32.

Chi, O.H., Jia, S., Li, Y., Gursoy, D., 2021. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. Comput. Hum. Behav. 118, 106700.

Childs, K. M. 2022. "The shade of it all": How black women use Instagram and youtube to contest colorism in the beauty industry. Soc. Media Soc., 8(2), 20563051221107634.

Cho, S., Mor Barak, M.E., 2008. Understanding of diversity and inclusion in a perceived homogeneous culture: A study of organizational commitment and job performance among Korean employees. Adm. Soc. Work 32 (4), 100–126.

Chu, E., Roy, D., Andreas, J. (2020). Are visual explanations useful? A case study in model-in-the-loop prediction. arXiv preprint arXiv:2007.12248.

Chuan, C.H., Tsai, W.H.S., Yang, J., 2023. Artificial Intelligence, advertising, and society. Advert. Soc. Q. 24 (3).

Colquitt, J.A., Scott, B.A., Judge, T.A., Shaw, J.C., 2006. Justice and personality: Using integrative theories to derive moderators of justice effects. Organ. Behav. Hum. Decis. Process. 100 (1), 110–127.

Curtin, N., Stewart, A.J., Cole, E.R., 2015. Challenging the status quo: The role of intersectional awareness in activism for social change and pro-social intergroup attitudes. Psychol. Women Q. 39 (4), 512–529.

Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V., Zou, J., 2021. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatol. 157 (11), 1362–1369.

Danks, D., London, A.J., 2017. Algorithmic bias in autonomous systems. Ijcai 17 (2017), 4691–4697.

Dastin, J., 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of Data and Analytics. Auerbach Publications, pp. 296–299.

Davis, B., Glenski, M., Sealy, W., Arendt, D., 2020. Measure utility, gain trust: practical advice for XAI researchers. In: 2020 IEEE Workshop on Trust and Expertise in Visual Analytics (TREX). IEEE, pp. 1–8.

De Cremer, D., De Schutter, L., 2021. How to use algorithmic decision-making to promote inclusiveness in organizations. AI Ethics 1 (4), 563–567.

Dierckx, K., Van Hiel, A., Valcke, B., van den Bos, K., 2023. Procedural fairness in ethnic-cultural decision-making: fostering social cohesion by incorporating minority and majority perspectives. Front. Psychol. 14, 1025153.

Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C., 2019. March). Explaining models: an empirical study of how explanations impact fairness judgment. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 275–285.

Downey, S.N., Van der Werff, L., Thomas, K.M., Plaut, V.C., 2015. The role of diversity practices and inclusion in promoting trust and employee engagement. J. Appl. Soc. Psychol. 45 (1), 35–44.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2012. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226.

Ehsan, U., Riedl, M.O., 2020. Human-centered explainable AI: Towards a reflective sociotechnical approach. In: HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22. Springer International Publishing, pp. 449–466.

Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., & Karahalios, K. (2019). User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-14.

Ford, C., Keane, M.T., 2022. Explaining classifications to non-experts: an XAI user study of post-hoc explanations for a classifier when people lack expertise. In: International Conference on Pattern Recognition. Springer Nature Switzerland, Cham, pp. 246–260.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun. ACM 64 (4), 136–143.

Giansiracusa, N., 2021. How Algorithms Create and Prevent Fake News. Apress, Berkeley, CA.

Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annu. Rev. Psychol. 62, 451–482.

Gunning, D., Aha, D., 2019. DARPA's explainable artificial intelligence (XAI) program. AI Mag. 40 (2), 44–58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z., 2019. XAI—Explainable artificial intelligence. Sci. Rob. 4 (37), eaay7120.

Hayes, A.F., 2017. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. Guilford Publications.

He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Hooker, S. 2021. Moving beyond "algorithmic bias is a data problem." Patterns, 2(4). Inuwa-Dutse, I. (2023). FATE in AI: Towards algorithmic inclusivity and accessibility. arXiv preprint arXiv:2301.01590.

Islam, S. R., Russell, I., Eberle, W., Dicheva, D. 2022. Incorporating the concepts of fairness and bias into an undergraduate computer science course to promote fair automated decision systems. Proceedings of the 53rd ACM Technical Symposium on Computer Science Education, 2, 1075-1075.

Johs, A.J., Agosto, D.E., Weber, R.O., 2022. Explainable artificial intelligence and social science: Further insights for qualitative investigation. Appl. AI Lett. 3 (1), e64.

Jones, D.A., Martens, M.L., 2009. The mediating role of overall fairness and the moderating role of trust certainty in justice–criteria relationships: The formation and use of fairness heuristics in the workplace. J. Organ. Behav. 30 (8), 1025–1051.

Jordan, S.L., Palmer, J.C., Daniels, S.R., Hochwarter, W.A., Perrewé, P.L., Ferris, G.R., 2022. Subjectivity in fairness perceptions: How heuristics and self-efficacy shape the fairness expectations and perceptions of organizational newcomers. Appl. Psychol. 71 (1), 103–128.

Jun, J., Kim, J.K., Woo, B., 2021. Fight the virus and fight the bias: Asian Americans' engagement in activism to combat anti-Asian COVID-19 racism. Race Just.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. Adv. Neural Inf. Proces. Syst. 33, 12104–12114.

Kenny, E.M., Ford, C., Quinn, M., Keane, M.T., 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. Artif. Intell. 294, 103459.

Kim, I., 2013. The relationship between critical ethnic awareness and racial discrimination: Multiple indirect effects of coping strategies among Asian Americans. J. Soc. Soc. Work Res. 4 (3), 261–277.

Kong, X., Liu, S., Zhu, L., 2024. Toward human-centered XAI in practice: A survey. Mach. Intell. Res. 1–31.

Kordzadeh, N., Ghasemaghaei, M., 2022. Algorithmic bias: Review, synthesis, and future research directions. Eur. J. Inf. Syst. 31 (3), 388–409.

Latu, I.M., Mast, M.S., Lammers, J., Bombari, D., 2013. Successful female leaders empower women's behavior in leadership tasks. J. Exp. Soc. Psychol. 49 (3), 444–448.

Ledford, H., 2019. Millions affected by racial bias in health-care algorithm. Nature 574 (31), 2.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., Mara, M., 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Comput. Hum. Behav. 139, 107539 https://doi.org/10.1016/j.chb.2022.107539.

Lin, C., Gao, Y., Ta, N., Li, K., Fu, H., 2023. Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. Telematics Inform. 85, 102068.

Lind, E.A., 2001. Fairness heuristic theory: Justice judgments as pivotal cognitions in organizational relations. Adv. Org. Just. 56 (8), 56–88.

Liu, B., 2021. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. J. Comput.-Mediat. Commun. 26 (6), 384–402.

Manresa-Yee, C., Roig-Maimó, M.F., Ramis, S., Mas-Sansó, R., 2021. Advances in XAI: Explanation interfaces in healthcare. In: Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects. Springer International Publishing, Cham, pp. 357–369.

McDermid, J.A., Jia, Y., Porter, Z., Habli, I., 2021. Artificial intelligence explainability: the technical and ethical dimensions. Phil. Trans. R. Soc. A 379 (2207), 20200363.

Melsión, G. I., Torre, I., Vidal, E., Leite, I. 2021, June. Using explainability to help children understand gender bias in AI. In Interaction Design and Children (pp. 87-99).

Mohseni, S., Zarei, N., Ragan, E.D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4), 1–45.

Montenegro, H., Silva, W., Gaudio, A., Fredrikson, M., Smailagic, A., Cardoso, J.S., 2022. Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy. IEEE Access 10, 28333–28347.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á., 2023. Human-in-the-loop machine learning: A state of the art. Artif. Intell. Rev. 56 (4), 3005–3054.

Murphy, M.C., Kroeper, K.M., Ozier, E.M., 2018. Prejudiced places: How contexts shape inequality and how policy can change them. Policy Insights Behav. Brain Sci. 5 (1), 66–74.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Seifert, C., 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Comput. Surv. 55 (13s), 1–42.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464), 447–453.

Parra, C.M., Gupta, M., Dennehy, D., 2021. Likelihood of questioning AI-based recommendations due to perceived racial/gender bias. IEEE Trans. Technol. Soc. 3 (1), 41–45.

Proudfoot, D., Lind, E.A., 2015. Fairness heuristic theory, the uncertainty management model, and fairness at work. The Oxford Handbook of Justice in the Workplace 1, 371–385.

Rai, A., 2020. Explainable AI: From black box to glass box. J. Acad. Mark. Sci. 48, 137–141.

Rice, D.B., Young, N.C., Sheridan, S., 2021. Improving employee emotional and behavioral investments through the trickle-down effect of organizational inclusiveness and the role of moral supervisors. J. Bus. Psychol. 36 (2), 267–282.

Rodrigues, R., 2020. Legal and human rights issues of AI: Gaps, challenges, and vulnerabilities. J. Respons. Technol. 4, 100005.

Rong, Y., Leemann, T., Nguyen, T.T., Fiedler, L., Qian, P., Unhelkar, V., Kasneci, E., 2023. Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Trans. Pattern Anal. Mach. Intell. 46 (4), 2104–2122.

Roovers, R. 2019. Transparency and responsibility in artificial intelligence. A call for explainable AI. Accessed on Apr 28, 2024.

Rothe, R., Timofte, R., Van Gool, L., 2015. Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10–15.

Russen, M., Dawson, M., Madera, J.M., 2021. Gender discrimination and perceived fairness in the promotion process of hotel employees. Int. J. Contemp. Hosp. Manag. 33 (1), 327–345.

Saasa, S.K., 2019. Discrimination, coping, and social exclusion among African immigrants in the United States: A moderation analysis. Soc. Work 64 (3), 198–206.

Saeed, W., Omlin, C., 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl.-Based Syst. 263, 110273.

Şahin, O., Van der Toorn, J., Jansen, W.S., Boezeman, E.J., Ellemers, N., 2019. Looking beyond our similarities: How perceived (in) visible dissimilarity relates to feelings of inclusion at work. Front. Psychol. 10, 425015.

Sanchez, J.I., Brock, P., 1996. Outcomes of perceived discrimination among Hispanic employees: is diversity management a luxury or a necessity? Acad. Manag. J. 39 (3), 704–719.

Schelenz, L., Bison, I., Busso, M., De Götzen, A., Gatica-Perez, D., Giunchiglia, F., Ruiz-Correa, S., 2021. The theory, practice, and ethical challenges of designing a diversity-aware platform for social relations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 905–915.

Schmitt, M., Dörfel, M., 1999. Procedural injustice at work, justice sensitivity, job satisfaction and psychosomatic well-being. Eur. J. Soc. Psychol. 29 (4), 443–453.

Schriber, S. 2023. Key skincare trends, https://civicscience.com/key-skincare-trends-mens-skincare-top-products-the-connection-with-mental-well-being/#:~:text=Skincare%20Routines%20and%20Comfortability%20Among%20Men%20and%20Women&text=Women%20are%20more%20than%20two,one%20daily%20(32%25%20vs (accessed on October 24, 2023).

Selbst, A., Powles, J. (2018, January). "Meaningful information" and the right to explanation. In: Conference on Fairness, Accountability and Transparency (pp. 48-48). PMLR.

Shen, H., DeVos, A., Eslami, M., Holstein, K. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. In *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1-29.

Shin, D., 2020a. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. Comput. Hum. Behav. 109, 106344.

Shin, D., 2020b. User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. J. Broadcast. Electron. Media 64 (4), 541–565.

Shin, D., 2021a. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int. J. Hum Comput Stud. 146, 102551.

Shin, D., 2021b. Why does explainability matter in news analytic systems? Proposing explainable analytic journalism. Journal. Stud. 22 (8), 1047–1065.

Shin, D., 2022. The perception of humanness in conversational journalism: An algorithmic information-processing perspective. New Media Soc. 24 (12), 2680–2704.

Shin, D., 2023a. Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. J. Inf. Sci. 49 (1), 18–31.

Shin, D.D., 2023b. Algorithms, Humans, and Interactions: How Do Algorithms Interact With People? Designing Meaningful AI Experiences, 1st ed. Routledge, New York.

Shin, D., Zhong, B., Biocca, F.A., 2020. Beyond user experience: What constitutes algorithmic experiences? Int. J. Inf. Manag. 52, 102061.

Shin, D., Kee, K.F., Shin, E.Y., 2022a. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? Int. J. Inf. Manag. 65, 102494.

Shin, D., Park, Y.J., 2019. Role of fairness, accountability, and transparency in algorithmic affordance. Comput. Hum. Behav. 98, 277–284.

Shin, D., Shin, E.Y., 2023. Data's impact on algorithmic bias. Computer 56 (6), 90–94.

Shin, D., Zaid, B., Biocca, F., Rasul, A., 2022b. In platforms we trust? Unlocking the black-box of news algorithms through interpretable AI. J. Broadcast. Electron. Media 66 (2), 235–256.

Shore, L.M., Chung, B.G., 2022. Inclusive leadership: How leaders sustain or discourage work group inclusion. Group Org. Manag. 47 (4), 723–754.

Song, H., Lu, H., McComas, K.A., 2021. The role of fairness in early characterization of new technologies: Effects on selective exposure and risk perception. Risk Anal. 41 (9), 1614–1629.

Starke, C., Baleis, J., Keller, B., Marcinkowski, F., 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data Soc. 9 (2).

Tanghe, S., 2016. Promoting critical racial awareness in teacher education in Korea: reflections on a racial discrimination simulation activity. Asia Pac. Educ. Rev. 17, 203–215.

Turner, J.C., 1987. Rediscovering the Social Group: A Self-Categorization Theory. Blackwell Publishing, Oxford, England.

Uzogara, E.E., Lee, H., Abdou, C.M., Jackson, J.S., 2014. A comparison of skin tone discrimination among African American men: 1995 and 2003. Psychol. Men Masculinity 15 (2), 201.

Van den Bos, K., Lind, E.A., 2004. Fairness heuristic theory is an empirical framework: A reply to Árnadóttir. Scand. J. Psychol. 45 (3), 265–268.

Van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M., 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. Artif. Intell. 291, 103404.

Van Prooijen, J.W., Van den Bos, K., Wilke, H.A., 2004. Group belongingness and procedural justice: Social inclusion and exclusion by peers affects the psychology of voice. J. Pers. Soc. Psychol. 87 (1), 66.

Wang, R., Harper, F. M., Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-14.

Wang, X., Yuen, K.F., Teo, C.C., Wong, Y.D., 2021. Online consumers' satisfaction in self-collection: Value co-creation from the service fairness perspective. Int. J. Electron. Commer. 25 (2), 230–260.

Zhou, J., Chen, F., Holzinger, A., 2020. Towards explainability for AI fairness. In: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers. Springer International Publishing, Cham, pp. 375–386.