Towards Effective Human Intervention in Algorithmic Decision-Making

Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions

Yurrita, Mireia; Verma, Himanshu; Balayn, Agathe; Gadiraju, Ujwal; Pont, Sylvia C.; Bozzon, Alessandro

DOI
10.1145/3706598.3713145

Publication date
2025

Document Version
Final published version

Published in
CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems

Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions

Mireia Yurrita
Delft University of Technology
Delft, Netherlands
m.yurritasemperena@tudelft.nl

Himanshu Verma
Delft University of Technology
Delft, Netherlands
H.Verma@tudelft.nl

Agathe Balayn
Delft University of Technology
Delft, Netherlands
a.m.a.balayn@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Sylvia C. Pont
Delft University of Technology
Delft, Netherlands
s.c.pont@tudelft.nl

Alessandro Bozzon
Delft University of Technology
Delft, Netherlands
a.bozzon@tudelft.nl

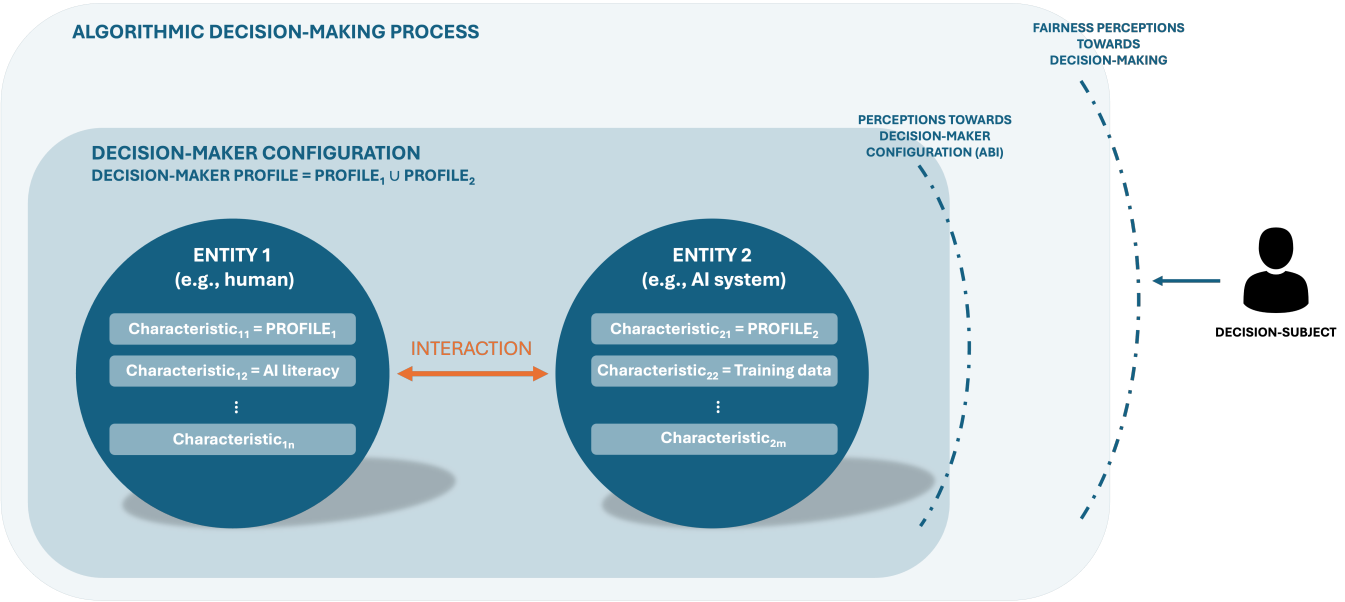


Figure 1: In our study, we evaluate the effect of different decision-maker configurations on decision-subjects' perceptions of decision-makers' ability, benevolence, and integrity. We also evaluate the relationship between decision subjects' perceptions of decision-makers and their fairness perceptions towards the algorithmic decision-making process. *Decision-maker configuration* refers to the collection of entities that compose a decision-making unit and the interactions among those entities. *Entity* refers to each independent element that composes a decision-maker configuration. *Characteristic* refers to the attributes that define the specificity of each entity. *Profile* refers to the characteristic that specifically describes the nature of each entity (e.g., human). The *union* of profiles that define the entities composing a decision-maker configuration constitutes the profile of the decision-maker configuration itself (e.g., if the profile of entity$_1$ is "human" and the profile of entity$_2$ is "AI system", the profile of the decision-maker configuration is "hybrid").

Abstract

Human intervention is claimed to safeguard decision-subjects' rights in algorithmic decision-making and contribute to their fairness perceptions. However, how decision-subjects perceive hybrid decision-maker configurations (i.e., combining humans and algorithms) is unclear. We address this gap through a mixed-methods

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713145

study in an algorithmic policy enforcement context. Through qualitative interviews (Study 1; N_1 = 21), we identify three characteristics (i.e., *decision-maker's profile*, *model type*, *input data provenance*) that affect how decision-subjects perceive decision-makers' ability, benevolence, and integrity (ABI). Through a quantitative study (Study 2; N_2 = 223), we then systematically evaluate the individual and combined effects of these characteristics on decision-subjects' perceptions towards decision-makers, and fairness perceptions. We found that only decision-maker's profile contributes to perceived ability, benevolence, and integrity. Interestingly, the effect of decision-maker's profile on fairness perceptions was *mediated* by perceived ability and integrity. Our findings have design implications for ensuring *effective* human intervention as a protection against harmful algorithmic decisions.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **Collaborative and social computing**; • **Computing methodologies → Machine learning**.

Keywords

human intervention, fairness perceptions, decision-maker, ability, benevolence, integrity

ACM Reference Format:

Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia C. Pont, and Alessandro Bozzon. 2025. Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3706598.3713145

1 Introduction

In the context of algorithmic decision-making, *human intervention* refers to the act of mediating an algorithmic output, where the (human) mediator has the appropriate competence and authority to potentially change this output [91]. Human intervention is included in regulatory efforts, like the European Union's General Data Protection Regulation (GDPR) [35], as a safeguard to protect decision-subjects'[1] "rights and freedoms and legitimate interests" against fully automated decisions[2]. By allowing a competent human to have control over automated decisions, *hybrid* decision-maker configurations (i.e., with human and artificial elements) are believed to offer the best of both worlds, i.e., the efficiency and data processing capabilities of Artificial Intelligence (AI) systems, and the flexibility of humans [38, 55, 85]. To evaluate the effectiveness of human intervention in algorithmic decision-making, the HCI community is increasingly examining the influence of different decision-maker configurations on decision-subjects' fairness perceptions (e.g., by varying the roles defined for humans and AI systems) [15, 66, 105]. Crafting algorithmic decision-making processes that uphold decision-subjects' standards of fairness is, in turn, key to ensuring the responsible implementation and broader acceptance [32, 60, 81, 100] of AI systems that could help deal with large-scale, increasingly-complex issues [38].

While previous HCI work capturing decision-subjects' fairness perceptions towards different decision-maker configurations has made important contributions, we identified two main research gaps. First, prior work mainly compared fully-automated configurations to exclusively-human configurations (e.g., [6, 20, 46, 60, 60, 64, 66, 82, 99]). In most cases, these studies concluded that decision-subjects prefer exclusively-human configurations [20, 60, 64]. Although these inquiries are valuable for understanding when algorithmic decision-making processes might not be desirable *at all*, they may not provide insights into whether and how humans can intervene in algorithmic processes to effectively safeguard decision-subjects against harmful automated decisions. The few studies that *did* compare fully-automated *vs.* hybrid decision-maker configurations found little evidence that confirms the effectiveness of human intervention in improving *decision-subjects'* fairness perceptions [103, 105]. Hence, informing the design of future algorithmic decision-making processes that appropriately integrate human input requires to further look into the effects of different *hybrid* decision-maker configurations on decision-subjects' fairness perceptions.

Second, prior work (e.g., [5, 6, 60, 93]) mainly evaluated characteristics of different decision-maker configurations (e.g., profile of the decision-maker, training data of the AI system, output explanations) *in isolation*, i.e., one characteristic at a time. However, prominent characteristics of such configurations might be intricately intertwined. For example, the role played by the AI training data in an algorithmic configuration depends on the decision-maker's profile. If the decision-maker is composed of (a) an AI system (i.e., fully-automated profile), it will rely on the training data to compute an output and make a decision; if the decision-maker is composed of (b) a combination of a human and AI system (i.e., hybrid decision-maker profile), the human will consider the AI output, which is conditioned by the training data, as an additional source of information —along with their knowledge and judgment— when making a decision. The perceived adequacy of the training data and the decision-maker profile might, therefore, co-shape decision-subjects' perceptions towards the decision-maker configuration and, jointly, impact decision-subjects' fairness perceptions. Identifying which decision-maker configuration is perceived as most beneficial by decision-subjects, therefore, requires *also* to look into the combined effects of diverse characteristics that define each configuration.

In this paper, we aim to inform ways in which humans can effectively intervene in *algorithmic* decision-making by capturing decision-subjects' (1) perceptions towards different decision-maker configurations and (2) fairness perceptions towards the decision-making process (see Figure 1). To this end, we adopted a mixed-methods approach [50] grounded in a context of algorithmic policy

[1]We will use the term *decision-subjects* to refer to individuals impacted by algorithmic decision-making.

[2]We will use the term *algorithmic* or *Artificial Intelligence (AI) system* to refer to computational systems for decision aid. We will use the term *algorithmic decision-making* to refer to decision-making processes that are driven or augmented by algorithmic systems —i.e., processes that are either *fully automated* or *hybrid*, respectively. To refer to decision-making processes where there is no algorithmic element, we will use the term *human decision-making*.

enforcement; specifically, the detection of illegal holiday rentals.[3] The mixed-methods approach consisted of two main stages:

(1) *Foundational interview study:* We first conducted interviews with 21 participants who rent their properties out for holiday purposes (Study 1; described in Section 3) —decision-subjects of illegal holiday rental detection. The interview study aimed to identify the characteristics that decision-subjects prioritize when assessing the adequacy of decision-maker configurations for this particular use case. The interview study also aimed to generate a preliminary understanding about how these characteristics might affect perceptions towards decision-makers' Ability, Benevolence and Integrity (ABI) [70]. We chose to characterize perceptions towards decision-makers through the ABI model [70] because this model distinguishes perceptions of trustworthiness towards decision-makers from trust (see section 2.3). The following research questions guided the interview study:
   - **RQ1.1.:** What are the main characteristics that decision-subjects consider when assessing the adequacy of decision-maker configurations?
   - **RQ1.2.:** How do these decision-maker characteristics relate to perceptions of ability, benevolence and integrity towards decision-makers?

   Through these qualitative interviews, we identified three prominent characteristics (i.e., decision-maker profile, model type, input data provenance) that affect decision-subjects' perceptions towards different decision-maker configurations (**RQ1.1.**). We mapped these characteristics onto the Ability, Benevolence, and Integrity (ABI) model [70] (**RQ1.2.**).
(2) *Large-scale quantitative study:* We then used the insights generated in the interviews to design a large-scale quantitative study (Study 2; described in Section 5). The objective of the large-scale quantitative study was to evaluate whether the preliminary insights generated in Study 1 are generalizable to a larger population and inform design decisions by decision-making entities. The following research questions guided our quantitative study:
   - **RQ2.1.**: How do characteristics related to decision-makers' configuration (i.e., *decision-maker profile*, *model type* and *input* data provenance) shape decision-subjects' perceptions of ability, benevolence, and integrity towards decision-makers?
   - **RQ2.2.**: How do perceptions of ability, benevolence, and integrity towards decision-makers predict decision-subjects' fairness perceptions towards algorithmic decision-making processes?

   For our quantitative approach, we designed an online, preregistered[4] user study. Participants were shown a scenario where a municipality would either incorporate a fully-automated or a hybrid decision-maker configuration to identify illegal holiday rentals. Decision-makers would make use of either a probabilistic or a rule-based model, fed with publicly or non-publicly available data. For each scenario, we measured perceived ability, benevolence, and integrity towards decision-makers and fairness perceptions towards the algorithmic decision-making process as a whole.

Our results show that the decision-maker profile (fully automated *vs.* hybrid) affected perceived ability and benevolence. Our exploratory analysis further shows that the decision-maker profile additionally may affect perceived integrity. In all cases perceptions towards hybrid decision-maker configurations were more favorable than fully-automated ones. We did not find a main effect of model type (rule-based *vs.* probabilistic) and data provenance (public *vs.* non-public) on perceived integrity. However, exploratory analyses suggest that there may be an interaction effect between the two characteristics (**RQ2.1.**). Our results also show that perceived ability and integrity positively relate to fairness perceptions (**RQ2.2.**). Furthermore, mediation analyses indicate that the effect of the decision-maker profile on fairness perceptions may be *mediated* by both perceived ability and integrity. In a similar vein, exploratory analyses suggest that the effect on fairness perceptions of participants' agreement with policy may also be mediated by perceived integrity. To ensure that human intervention safeguards decision-subjects' rights, freedoms, and legitimate interests, our findings encourage public agencies implementing algorithmic decision-making processes to (a) design workflows where street-level bureaucrats can effectively intervene, (b) balance the need for justifying algorithmic decisions with decision-subjects' right to privacy, (c) disentangle perceptions towards decision-makers and the implemented policy, and (d) engage with impacted communities when designing human intervention. Our findings additionally encourage future HCI research to (e) further examine the effectiveness of hybrid decision-maker configurations in real-world contexts and (f) account for the complex and distributed human labor that AI systems result from.

In this paper we, therefore, make two main contributions.

- We generate empirical data on the individual and combined effects of decision-maker profile, model type and input data provenance on perceptions of ability, benevolence and integrity, and we identify how these perceptions relate to fairness perceptions.
- Drawing from those empirical insights, we provide four recommendations for public agencies developing and deploying AI systems for decision-making.

2 Related Work

This section first introduces the concept of *human intervention* for algorithmic decision-making (section 2.1). We then summarize recent research looking into fairness perceptions towards human intervention in algorithmic decision-making (section 2.2). We finally give an overview of different models capturing perceptions towards decision-makers (i.e., models of trust and perceived trustworthiness) and their relation to fairness perceptions (section 2.3).

2.1 Human Intervention in Algorithmic Decision-Making

In the context of algorithmic decision-making, *human intervention* is defined by regulatory efforts, such as the European Union's General Data Protection Regulation (GDPR) [35], as the act of providing human input by an individual with the competence and

[3] We chose an algorithmic system suggested by the municipality of Amsterdam for detecting illegal holiday rentals as a use case. https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/ (last accessed 11.09.2024)

[4] The preregistration is available at https://osf.io/82c95 (preregistered on 10.12.2023)

authority to change an algorithmic output [91]. In Article 22(3) of the GDPR [35], human intervention is represented as one of the three measures —along with decision-subjects' right to express their point of view and to contest automated decisions— that safeguard decision-subjects' "rights, freedoms, and legitimate interests" against fully-automated decision-maker configurations. Legal scholars have framed human intervention as a means to protect decision-subjects' fundamental right of human dignity [4]; a measure to acknowledge the "foundational indeterminacy of human self" [48, 73]. By allowing a competent human to provide input, human intervention is also claimed to be "an antidote to machine error" [4].

The role of human intervention is especially important in the public sector. Public decision-making processes deal with societally-sensitive topics [92, 107], where decision-subjects do not have an alternative to dealing with public administration [2] —unlike the private sector, where decision-subjects can stop using a service if they are not satisfied with it. Decision-making processes in the public sector rely on the interpretation of policy performed by *street-level bureaucrats* (i.e., civil servants that directly interact with citizens) [3, 107]. At the decision-making time, street-level bureaucrats engage in *reflexivity* [3] and account for decision-subjects' individual circumstances for turning *defined* policies into *effective* policies, i.e., they apply *administrative discretion*. When AI systems are introduced for public decision-making, decisions are made based on decision-subjects' position with respect to the algorithm's decision boundary. Any corrective feedback to consider decision-subjects' individual circumstances is gathered and applied *after* decision-making [3]. Human intervention in the public sector aims at retaining and restoring street-level bureaucrats' discretionary power as part of algorithmic decision-making [107]. Given the relevance of human intervention in algorithmic decision-making in the public sector, we ground our study in a policy enforcement context.

2.2 Decision-Subjects' Fairness Perceptions Towards Decision-Maker Configurations in Algorithmic Decision-Making

In an effort to test the effectiveness of human intervention in protecting decision-subjects' rights in algorithmic decision-making, the number of HCI studies capturing decision-subjects' fairness perceptions towards various decision-maker configurations has proliferated [96, 101].[5]

A considerable amount of work has been devoted to comparing fairness perceptions towards human *vs.* fully-automated decision-makers [6, 20, 29, 46, 57, 60, 63, 78, 82]. Most prior work has claimed that people normally considered humans to be more fair [20, 29, 46, 57, 60, 65, 78, 82]. Preference towards humans has been claimed to be caused by the perceived facility to convince them towards a favorable outcome as compared to algorithmic systems [40] and humans' ability to account for non-quantifiable aspects of the decision-making [78]. While comparing fully-automated *vs.* exclusively-human decision-maker configurations is valuable to determine cases where algorithmic decision-making might not be desirable *at all*, it generates little insight into whether and how humans can intervene in algorithmic decision-making.

A smaller number of studies [77, 103, 105] has compared fully-automated *vs.* hybrid (i.e., involving humans and algorithmic systems) decision-makers. These studies have, counterintuitively, led to inconclusive results. On the one hand, Wang et al. [103] and Yurrita et al. [105] did not find any significant differences between both profiles. In both cases, the hybrid decision-maker configuration consisted of a human who would supervise every algorithmic decision [103] or those cases where the confidence of the AI output was low [105] (i.e., the interaction between the human and AI was based on *supervisory control* [89]). Nagtegaal [77], on the other hand, found that procedural justice perceptions could increase when the decision was made by a hybrid decision-maker for low-complexity tasks. In this case, the interaction in the hybrid decision-maker configuration was based on *advisory control* [89], where the human would evaluate the output given by the AI system. However, the preference towards hybrid decision-makers was only true for high-complexity tasks if both options (hybrid and human decision-maker) were juxtaposed through a within-subject setup but did not hold if the setup was between subjects [77]. Motivated by the absence of conclusive evidence, this paper aims to deepen the understanding of the impact of human intervention in algorithmic decision-making. To this end, we evaluate decision-subjects' perceptions towards different decision-maker configurations that include algorithmic elements and varying levels of human input.

Recent work has also tested the effect of additional decision-maker-related characteristics on decision-subjects' fairness perceptions. These studies (mainly) tested the effect of one characteristic at a time. The most prominent ones are explanations [15, 30, 93, 105], the decision basis [41], details about the design [103] and data to train the system [5]. Explanations have been found to have a positive effect on informational fairness perceptions [93, 105], which, in turn, positively relate to overall fairness perceptions [105]. Using features that are perceived as relevant as the decision basis has been found to lead to positive fairness perceptions [41]. No evidence has been found of development procedures (e.g., developed in-house *vs.* outsourced) affecting fairness perceptions [103]. Information about the data used to train the system has been found to help users assess the fairness of a system [5].

To account for the potential entanglements between several decision-maker-related characteristics, in our study, we systematically evaluate the individual *and* combined effects of prominent decision-maker-related characteristics.

2.3 Models Capturing Perceptions Towards Decision-Maker Configurations and Their Relation to Fairness Perceptions

Fairness perceptions towards algorithmic decision-making have been captured in various different ways. A recent systematic review by Starke et al. [96] showed that fourteen of the reviewed studies directly captured fairness perceptions through single items. Instead, seventeen of the reviewed studies used fairness scales

[5]Note that we refer to literature that captured decision-subjects' perceptions towards different decision-maker configurations. Our related work section, therefore, does not include studies about the effect of different algorithmic configurations on end-users' trust/reliance or studies optimizing AI systems for teamwork in hybrid decision-maker configurations (e.g., [10, 106]).

designed for human decision-making and adapted them to algorithmic decision-making. One of the most popular fairness scales is the one suggested by Colquitt [24]. Colquitt [24] defined fairness perceptions across four justice dimensions: *distributive* (i.e., dimension related to decision outcomes), *procedural* (i.e., related to the process), *interpersonal* (i.e., related to the treatment towards decision-subjects) and *informational* (i.e., related to the provided information). Despite the widespread usage of Colquitt [24]'s scale, the suggested dimensions put little emphasis on evaluating the adequacy of the decision-maker configuration. The interpersonal justice dimension, for instance, captures whether decision-subjects were treated with respect during their interaction with decision-makers. However, it does not capture decision-subjects' perceptions towards the decision-maker configuration itself. This might make it difficult to disentangle potential reasons why decision-subjects might deem the decision-maker configuration (in)appropriate [70].

In organizational psychology, methods for capturing perceptions towards decision-maker configurations have instead been characterized as models of trust (e.g., [26, 70, 80]). Some scholars [26, 62, 71, 80, 86, 87] conceptualize trust as the trustor's (i.e., party that trusts another party) positive expectations towards the trustee's (i.e., party that is trusted) conduct, motives, and intentions in a situation that entails risk. This generates in the trustor a willingness to act based on the trustee's words, actions or decisions [25]. An alternative line of work has studied trust as the trustor's willingness to be vulnerable to the trustee's actions [69, 70]. Mayer et al. [70]'s work is especially influential in this research area. Mayer et al. [70] define ability, benevolence and integrity (i.e., ABI model) as factors contributing to the perceived trustworthiness of the trustee. *Ability* refers to a set of competencies or skills that the trustee possesses and that enable the trustee to influence the decision-making domain [70]. *Benevolence* refers to the goodwill of the trustee towards the trustor [70]. *Integrity* is defined as the trustor's perception that the trustee adheres to an acceptable set of principles [70]. Trust is conceptualized as a result of the trustee's perceived trustworthiness, along with the trustor's propensity to trust in a risk situation. Previous work on automation has built on Mayer et al. [70]'s model and adapted it to scenarios where the trustee is an automated agent [14, 56, 72].

In this paper, we inform ways for humans to effectively intervene in algorithmic decision-making by first capturing perceptions towards different decision-maker configurations. To this end, we follow Colquitt and Rodell [25] and adopt the ABI model [70] to characterize perceptions towards decision-makers. The reason for adopting the ABI model [70] and not other trust models [26, 62, 71, 80, 86, 87] is that the ABI model [70] distinguishes perceptions of trustworthiness towards decision-makers from trust. The ABI model [70] characterizes perceived trustworthiness as an antecedent to trust and captures it separate from trustor-related factors (e.g., propensity to trust) or contextual factors (e.g., perceived risk). This distinction between trust and trustworthiness can bring conceptual clarity and precision to capture perceptions towards decision-maker configurations (conceptualized as their ability, benevolence, and integrity) and evaluate their effect on fairness perceptions [25]. We apply the ABI model [70] in its original form. While studies in automation have shed light on how to adapt the ABI model [70] to automated decision-making scenarios, their focus has been on capturing *end-users'* (i.e., individuals interacting with the automated system) perceived trustworthiness towards the *automated system* (e.g., e-commerce agents [14], AI-enabled technology [42], AI for decision aid [94]). In our study, however, we focus on *decision-subjects'* (i.e., individuals impacted by the decision-making process) perceptions towards *decision-maker configurations* and their effect on fairness perceptions. To the best of our knowledge, Höddinghaus et al. [49]'s work has been the only one that characterized decision-makers from a decision-subject perspective for *algorithmic* decision-making. Höddinghaus et al. [49] characterized decision-makers through the original ABI model [70] and adjusted *ability* items to capture two relevant facets of algorithmic decision-making: data processing capacity and adaptability to changing conditions. We follow Höddinghaus et al. [49]'s approach and apply the original ABI model [49] with adapted ability items to capture perceptions towards decision-maker configurations. Unlike Höddinghaus et al. [49], we use this approach to compare decision-subjects' perceptions towards fully-automated *vs.* hybrid decision-maker configurations.

Perceptions towards decision-maker configurations and fairness perceptions towards the decision-making process are, in turn, highly connected. Several theoretical works (e.g., [97, 98]) have noted the existence of relationships between perceived trustworthiness and fairness perceptions. In an empirical study, Colquitt and Rodell [25] showed that the relationship between perceived trustworthiness towards decision-makers —conceptualized through the ABI model [70]— and fairness perceptions is reciprocal for *human* decision-making. To the best of our knowledge, no prior work has investigated how trustworthiness perceptions towards decision-maker configurations conceptualized as the decision-makers' ability, benevolence, and integrity affect fairness perceptions in *algorithmic* decision-making. We do so in this study. On a practical level, we believe that evaluating this relation can inform design decisions by decision-making entities. For example, if, based on the decision-maker configuration, decision-subjects have already formed negative perceptions of ability, and this strongly affects decision-subjects' fairness perceptions, making changes in appeal mechanisms –element beyond the decision-maker configuration that has been shown to contribute to fairness perceptions [105]– might not be effective; changes in the decision-maker configuration itself should be prioritized. On an empirical level, it also allows us to bring nuance to the relation between trustworthiness and fairness constructs, and capture whether and how fairness perceptions relate differently to each of ABI [70] dimensions.

3 Study 1: Foundational Interview Study

In this paper, we adopt a *mixed-methods approach* [50]. We followed prior work [12], and first conducted a foundational interview study (1) to identify the main characteristics that participants highlighted when evaluating the adequacy of decision-maker configurations for an illegal holiday rental detection scenario and (2) to get a preliminary understanding on how these might relate to perceptions towards decision-makers' ability, benevolence, and integrity. In contrast to [12], we focused on decision-subjects' perceptions towards decision-maker configurations, rather than perceptions of industry experts towards AI systems. We, then, used these findings

to formulate our research questions, hypotheses, and to design our quantitative study (as described in Sections 4 and 5). All supplementary materials linked to this paper can be found in our repository (https://osf.io/r8whs/). These include the interview protocol and prompts used for the qualitative study and the preregistration, task design, data, and code for analysis of our quantitative study.

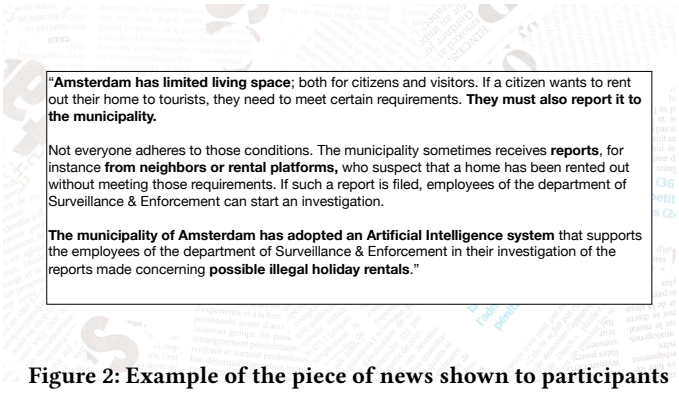
3.1 Use Case and Participant Recruitment

3.1.1 Illegal Holiday Rental Detection. For our study, we focused on illegal holiday detection as a *use case* within the context of algorithmic policy enforcement. In recent years, the proliferation of short-term rentals (e.g., Airbnb) in highly populated cities has led to municipalities increasing their efforts to regulate those rentals (e.g., Amsterdam, Barcelona), or, in some cases to ban some listings (e.g., New York City) [13, 79]. To identify illegal holiday rentals and address the presented issue, municipalities all over the world[6] have suggested workflows to search for potential illegal holiday rentals. For Study 1, we chose to focus on the algorithmic system suggested by the municipality of Amsterdam[7] to identify illegal short-term rentals. The municipality of Amsterdam developed a risk-based system that prioritizes reports submitted by citizens by relying on features about the identity of the reported property owner, building data, and prior illegal housing cases. This system was suggested in 2019 and expected to be pilot tested in 2020.

Although this system has, to date, not been deployed due to delays in data collection caused by the COVID-19 pandemic,[8] we argue that this use case represents a compelling context for our study. The reasons for this are threefold. (1) It is a timely decision-making process that deals with a widespread issue and for which algorithmic systems might be used in the near future. (2) It is a real-world use case and, therefore, allows us to inform municipalities on the design of algorithmic systems that are aligned with decision-subjects' fairness perceptions. (3) It also allows us to recruit participants that could potentially be affected by similar systems in the future.

3.1.2 Participant Recruitment. We recruited 21 participants from Western countries with experience renting their properties out as short-term rentals and that could potentially be correctly or incorrectly identified by these types of systems (i.e., they had a personal stake in the topic [21], and, therefore, represented proxy decision-subjects). Since the topic at hand affects a wide range of highly populated cities in several Western countries, we decided not to limit the study to the Amsterdam area and included participants who rent out properties in cities where initiatives to identify illegal holiday rentals (algorithmic or not) have been put in place. We also ensured diversity in participants' disciplinary backgrounds and self-reported AI literacy —see Table 1. We recruited participants by announcing our study in our institution and in short-term rental channels, and by reaching personal contacts.

3.1.3 Interview Procedure. In line with previous research (e.g., [16, 34, 52]), we used a scenario-based approach to introduce our participants to the use case.[9] We introduced a fictional piece of news describing the use case (see Figure 2) and we asked our participants about their perceptions towards the benefits and drawbacks of introducing an AI system for the detection of illegal holiday rentals. We additionally showed our participants the information about the system as summarized in the algorithm register (e.g., data provenance, type of algorithm, workflow, potential harms). This allowed us to obtain a nuanced understanding of the aspects of the system and the decision-maker configuration that participants perceived as (in)appropriate. Note that participants were not directly asked about their perceptions of ability, benevolence, and integrity towards the decision-maker configuration. These connections were drawn as a result of the analysis process.


"**Amsterdam has limited living space**; both for citizens and visitors. If a citizen wants to rent out their home to tourists, they need to meet certain requirements. **They must also report it to the municipality.**

Not everyone adheres to those conditions. The municipality sometimes receives **reports**, for instance **from neighbors or rental platforms,** who suspect that a home has been rented out without meeting those requirements. If such a report is filed, employees of the department of Surveillance & Enforcement can start an investigation.

The municipality of Amsterdam has adopted an Artificial Intelligence system that supports the employees of the department of Surveillance & Enforcement in their investigation of the reports made concerning **possible illegal holiday rentals**."

Figure 2: Example of the piece of news shown to participants to introduce our use case. The material used for each participant included the name where their short-term rental was located.

3.1.4 Data Collection and Analysis. We conducted one-hour online interviews between July and August 2023. Before conducting our study, our research plan was reviewed and approved by the ethics committee in our institution. The participation in our study was compensated with 25 EUR or equivalent in local currency. The recordings of the interviews were transcribed and analyzed using *thematic analysis* [21, 22] with a combination of inductive and deductive orientation to data. The analysis process took place in an iterative way, moving between empirical data and theory. The first author inductively explored the empirical data and generated a first set of codes. The second and third authors partially coded the data. We then consulted Mayer et al. [70]'s model and deductively grouped the codes into the dimensions of ability, benevolence and integrity. While these dimensions might overlap at times (e.g., a decision-maker showing empathy could be considered to have the

[6] See the examples of New York City: https://portal.311.nyc.gov/article/?kanumber=KA-02317; Barcelona: https://meet.barcelona.cat/habitatgesturistics/en; Berlin: https://ssl.stadtentwicklung.berlin.de/wohnen/zweckentfremdung_wohnraum/formular/adresswahl.shtml; or Porto: https://www.asae.gov.pt/espaco-publico/formularios/queixas-e-denuncias.aspx

[7] https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/(last accessed 11.09.2024)

[8] Check the official communication on the status of the project https://amsterdam.raadsinformatie.nl/document/12731876/2#search=%22Afhandeling%20toezegging%20pilot%20algoritme%20Alpha%20handhaving%20vakantieverhuur%22 (last accessed 11.09.2024)

[9] In this same interview, we also inquired participants about their needs for contesting this algorithmic decision-making process. We have used that data for another publication.

Table 1: Summary of our 21 interviewees' demographics. Note that two of our participants have a joint background in Business and Law.

Feature	Category (Number of participants)
Self-reported AI literacy	High (7), Medium (7), Low (7)
Background	Computer Science (5), Engineering (4), Law (4), Business (3), Design (3), Architecture (2), Physics (1), Social Work (1)
Country	Netherlands (9), Spain (7), US (2), Portugal (1), Germany (1), Canada (1)
Immigration status	Native (12), Non native (9)

ability to be empathetic —*ability* dimension— or having the willingness to do good —*benevolence* dimension—), we identified the strongest association between the code groups that we generated in the analysis and Mayer et al. [70]'s definition of each dimension (e.g., we interpret empathy as "a positive orientation of the trustee towards the trustor" [70] even when there is no extrinsic reward, and cluster it within the dimensions of perceived benevolence). We then reflected on the characteristics of the decision-maker configuration relative to which participants were evaluating the adequacy of the configuration, i.e., the characteristics that might cause the observed variations in perceptions. In most cases, participants would not explicitly mention the characteristic that caused variations in their perceptions, but the identification of such characteristics was the result of the interpretative process that the authors engaged in [21, 22] –see Figure 3. Unlike [12], in this paper we do not intend to provide an exhaustive set of *all* characteristics that might affect participants' perceptions but rather identify a set of characteristics whose effect we can then quantitatively test. We narrowed the number of characteristics down applying two main criteria: (1) the characteristic was prominent and (2) the total number of characteristics was tractable quantitatively. We, therefore, report a list of three characteristics that caused variations in perceptions of at least one third of participants. Note that the extent to which characteristics deemed prominent in the interviews were, indeed, relevant to a larger population was then quantitatively tested through Study 2.

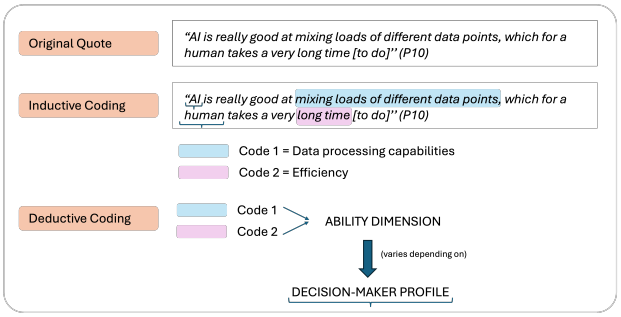


Figure 3: Example of the analysis process for one quote.

3.2 Findings From Study 1

In the following lines, we map prominent **characteristics** relative to which our participants evaluated the adequacy of decision-makers to each dimension of Mayer et al. [70]'s model (i.e., ability, benevolence, integrity).

3.2.1 ***Perceived Ability.*** Overall, participants were optimistic about integrating an AI system into the decision-making process. 19 out of 21 participants evaluated the ability of the decision-maker to detect illegal holiday rentals based on the **decision-maker's profile**, from fully-automated decision-makers to hybrid decision-maker configurations, i.e., with the intervention of civil servants (*"We have been using AI to deal with data since a long time ago. It depends on which level of autonomy the AI has."* P6). When referring to the necessary competencies, and characteristics of the decision-maker, most participants highlighted accuracy as one of the most important dimensions. Many (13/21) pointed out the data processing capabilities of AI systems, considering AI systems effective tools for initial screening (*"I suppose you could design an AI system that would flag questionable complaints that, you know, need to be investigated in some way."* P7). AI systems were believed to be able to detect patterns that humans cannot (9/21). Efficiency was considered the main reason to implement an AI system (12/21), viewing it as a good way of dealing with bureaucracy.

Even if AI systems were seen as a means to improve decision accuracy, several participants acknowledged the imperfect nature of AI (7/21) and the importance of ensuring good quality input data (4/21) (*"I think a human has to be behind it. I would use the AI to flag the ones [reported properties], and rank the ones that could be more illegal. But, sometimes there can be errors, or some houses maybe have an old license. I know that databases can be outdated. There has to be someone checking."* P10). Remarks about AI (in)accuracy and (lack of) data quality were often made to highlight that decision-maker configurations should include some level of human intervention *at decision-making time*: civil servants were seen as capable of correcting errors made by the AI system during the interaction (8/21). Only a few (3/21), were interested in knowing about human intervention in the definition of training data or during AI system development for evaluating the ability of the decision-maker configuration.

3.2.2 ***Perceived Benevolence.*** 13 out of 21 participants evaluated the decision-maker's willingness to do good (i.e., *benevolence* [70]) based on the **decision-maker's profile**. AI systems were seen as unable to account for contextual factors and to allow decision-subjects to discuss and argue, which is needed to treat decision-subjects with consideration (6/21) (*"They [human civil servants] need to use their more personal human skills. Maybe they [owners] can lie, but, you still give the owner a chance to at least defend and argue."* P10). Civil servants, instead, were considered to be willing to understand the "shades" of the decision-making process, and to offer a full picture of the situation to a partial AI (8/21); if

civil servants made the last decision, decision-subjects would not be reduced to numerical values. A few (3/21) highlighted that civil servants should show empathy and politeness towards the decision-subject (*"I would prefer to have the point of view of a person that can also really understand me. A real person who is available to explain, who is polite, who is available to give information. And to help me also."* P12). Others (5/21) additionally mentioned care, commitment, and consideration as necessary properties for decision-makers to be considered benevolent (*"I [as a decision-subject] want to talk to someone that can understand what I'm afraid of and not to someone that will tell me on the phone: yeah, this is not right."* P18).

3.2.3 ***Perceived Integrity.*** 20 out of 21 participants evaluated the decision-makers' integrity based on the means that these use for making the decision, namely the decision basis operationalized as the **model type** (i.e., probabilistic *vs.* rule-based). For ensuring integrity, those participants highlighted that the decision basis should comprise relevant and actionable features, where the cause of the decision should be clearly stated in relation to the rules violated by the decision-subject (*"But where is the proof that it [illegal rental] is so? That I have a 35 m2 apartment? And that a neighbor has called to complain about that? It proves that I am renting my home illegally? I don't think so, if this is not backed up with other data."* P9). 9 out of 21 participants evaluated the decision-makers' integrity based on input **data provenance** (i.e., publicly *vs.* non-publicly available). Those participants indicated that the information used for decision-making should be aligned with the principle of proportionality, i.e., come from an ethically acceptable source (*"If they have a movie or camera, a picture with a large group of people, people moving in the house with big backpacks. In that case, I would question if they are using the data for the purpose that the data was generated."* P20). Facilitating fraud detection was seen as positive to avoid a shortage of long-term rentals, which was seen as a social good (4/21) (*"There might be many citizens who don't have access to housing, and I believe housing is a human right. So if this algorithm is being used to identify cases where the house that is being rented should be given to citizens instead of tourists. Then I think this AI is doing something good."* P4).

4 Hypotheses For Study 2

Combining the insights we got from our qualitative study with prior literature in algorithmic and human decision-making (e.g., [15, 25, 93, 103]), we formulated seven hypotheses about the effect of characteristics defining a decision-maker configuration on perceived ability, benevolence, and integrity, and the effect of these on fairness perceptions. An overview of the hypotheses is given in Figure 4. All seven hypotheses were pre-registered before collecting the data. The combined effects between characteristics were examined in an exploratory fashion (see section 6.3).

4.1 Hypotheses Related to RQ1: Characteristics Affecting Perceived Ability, Benevolence, Integrity

- ***Hypothesis 1a (H_{1a}).*** **A hybrid decision-maker configuration (i.e., with human intervention)[10] is perceived as more able than a fully-automated one.**
  Rationale. In Study 1, we observed that 19 out of 21 participants evaluated decision-makers' ability based on the decision-maker profile, hybrid configurations being considered as the ones that bring the best of the AI system and the human. Previous work suggests that fully-automated decision-maker configurations are perceived to be efficient and objective [60, 105]. However, these are also perceived to be less adaptable than humans [49]. Participants in our qualitative study highlighted that a hybrid decision-maker benefits from the ability of the algorithmic system to efficiently and accurately process data, while enabling the human to exercise discretion. We, therefore, hypothesize that a hybrid decision-maker configuration will be perceived as more *able* than a fully-automated decision-maker configuration.
- ***Hypothesis 1b (H_{1b}).*** **A hybrid decision-maker configuration is perceived as more benevolent than a fully-automated one.**
  Rationale. In Study 1, we observed that 13 out of 21 participants evaluated decision-makers' benevolence based on the decision-maker profile, configurations relying only on AI systems being considered as unemphatic and rigid. Previous work, through qualitative findings, also suggests that fully-automated decision-maker configurations are considered impersonal and dehumanizing [15]. Problematic aspects of fully-automated decision-maker configurations include their inability to account for the unique individual circumstances of decision-subjects, and to adapt the decision-making to their needs and preferences [64, 105]. In our qualitative study, participants highlighted that a decision-maker where the final decision is made by a human, can show empathy and consideration towards the decision-subject, i.e., can be more benevolent. We, therefore, hypothesize that a hybrid decision-maker configuration will be perceived to be more *benevolent* than a fully-automated algorithmic decision-maker configuration.
- ***Hypothesis 1c (H_{1c}).*** **The perceived integrity of a decision-maker configuration is higher when it concerns rule-based models than when it concerns a probabilistic model.**
  Rationale. In Study 1, we observed that 20 out of 21 participants assessed decision-makers' integrity based on the model type. Binns et al. [15], through their qualitative findings, suggested that decision-subjects consider statistical inferences unacceptable as a basis for algorithmic decision-making. Similarly, some participants of our qualitative study claimed that generalization should not be acceptable as a decision basis, and that decisions should not be supported by a system that relies on what other individuals did. Participants, in contrast, were asking for a clear indication of the rules that they were violating. Even if Wang

[10] In the pre-registration, we formulated our hypotheses by referring to hybrid decision-maker configurations as "a human decision-maker that uses an algorithmic system to augment their capabilities" and fully-automated decision-makers as "algorithmic decision-makers". For the sake of consistency with the rest of the paper, we will use the term "hybrid decision-maker configuration" *vs.* "fully-automated decision-maker configuration".

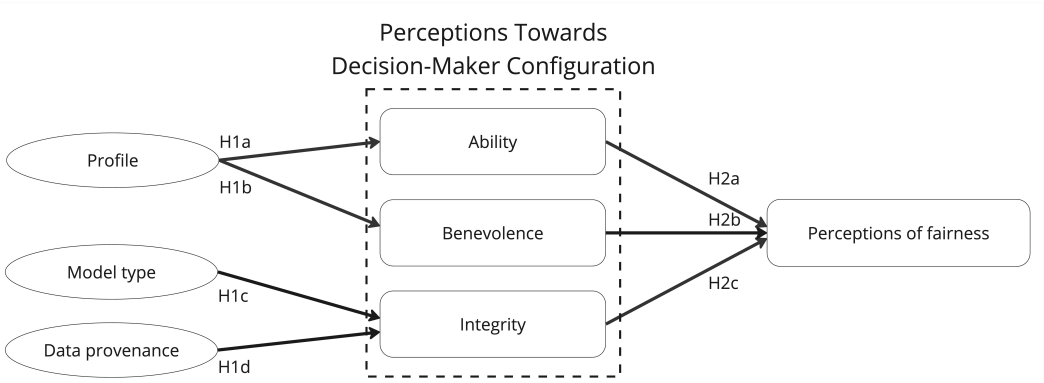


Figure 4: Overview of our hypotheses.

et al. [103] did not find any effect of the model type on decision-subjects' fairness perceptions, we hypothesize that relying on rule-based models will contribute to higher perceptions of *integrity* compared to probabilistic models.

- ***Hypothesis 1d (H_{1d}).* The perceived integrity of a decision-maker configuration is higher when the data used for decision-making comes from publicly available databases rather than non-publicly available data sources.**
  Rationale. In Study 1, we observed that 9 out of 21 participants assessed decision-makers' integrity based on the input data provenance. These participants suggested that it is acceptable to use publicly available data as input data while accessing data that might invade the privacy of decision-subjects (i.e., non-publicly available data) was not considered acceptable. Previous work showed that information about data sources used for training a model allows users to judge the trustworthiness of a system and to assess its fairness [5]. Even if the effect found by Anik and Bunt [5] referred to training data rather than input data, we hypothesize that the type of input data will affect decision-subjects' perceptions. More concretely, using non-publicly available data for decision-making will negatively impact decision-subjects' perceptions of *integrity* towards the decision-maker as compared to using publicly available data.

4.2 Hypotheses Related to RQ2: Effect of Perceived Ability, Benevolence, Integrity on Fairness Perceptions

- ***Hypothesis 2a (H_{2a}).* Perceived ability relates positively to perceptions of fairness.**
  Rationale. Previous literature in human decision-making did not find ability to be a significant predictor for fairness perceptions [25]. As opposed to these findings, we hypothesize that a difference in context might play a role. Colquitt and Rodell [25] studied the relationship between perceived ability and perceptions of fairness by recruiting alumni from a university and capturing their perceptions towards their immediate managers. For this context, the authors argued that more able managers might create more outcome differentiation in their units, which the alumni might not always benefit from, and therefore, might not perceive as fair. As opposed to this context, we hypothesize that in a context where citizens might benefit from higher levels of ability in the decision-maker (e.g., by ensuring that, thanks to detecting illegal holiday rentals, the societal issue of not having enough long-term rental availability is ameliorated), perceived ability will relate positively to fairness perceptions.
- ***Hypothesis 2b (H_{2b}).* Perceived benevolence relates positively to perceptions of fairness.**
  Rationale. Prior literature in human decision-making found that for benevolence, the relationships between perceived trustworthiness and fairness perceptions are reciprocal; both influencing one another [25]. Similarly, we hypothesize that in *algorithmic* decision-making, benevolence will relate positively to fairness perceptions.
- ***Hypothesis 2c (H_{2c}).* Perceived integrity relates positively to perceptions of fairness.**
  Rationale. Literature in human decision-making has shown that perceptions of *integrity* affect dimensions of distributive, procedural, informational and interpersonal fairness perceptions [25]. We hypothesize that for *algorithmic* decision-making processes, there will also be a positive relation between perceived integrity and fairness perceptions.

5 Study 2: Large-Scale Quantitative Study

In this section, we describe how the insights generated in Study 1 (section 3) informed the design of our quantitative study. Our quantitative study aims at testing the hypotheses (see section 4) formulated based on the understanding we gained through the interview-based study.

5.1 Variables

5.1.1 Independent Variables. To capture perceptions towards decision-makers while avoiding *outcome favorability bias* [65, 103], the scenario shown to our participants was narrated in the third person and we asked them to look into it through the lens of a decision-subject, following prior work [6, 93, 95, 105]. We generated $2 \times 2 \times 2 = 8$ different scenarios based on three independent variables.

Table 2: Overview of independent variables and their origin.

Independent Variable	Conditions	Origin
Profile	Hybrid	Examples given by participants in Study 1 (e.g., *"So, if it's just something that is supporting the human decision-making when dealing with huge amounts of data, I think that's fine."* (P6)).
	Fully-automated	Previous work where AI makes the final decision [103].
Model type	Probabilistic	Original system designed by Amsterdam municipality. Unlike prior work [103], we did *not* use terms like "machine learning" to refer to probabilistic models to make the provided information accessible to participants with all levels of AI literacy and to avoid *ambiguity bias* (i.e., association of negative perceptions to missing or ambiguous information [31]).
	Rule-based	20 out of 21 participants' desire to be evaluated in relation to the rules they had violated in Study 1.
Data provenance	Publicly available databases	Workings of the original system suggested by the municipality of Amsterdam.
	Non publicly available data sources	Examples given by participants in Study 1 (e.g., *"You could use street cameras to determine how many people stay there for which period of time"* (P21)).

- **Profile** (*categorical, between-subjects*). Each participant was randomly assigned to one of two configurations (Table 2):

(1) Hybrid (AI-Human). An AI was used as a screening tool that informs the decision of the human civil servant to consider the reported property an illegal holiday rental. The human civil servant would evaluate the output of the system and, based on their own judgment [77], decide whether to send a first warning to the property owner.[11]

(2) Fully-automated (only AI). An AI would evaluate the reported property and, based on that evaluation, determine whether there is an illegal holiday rental in that address. Based on the output of the AI system, a warning letter would be sent to the property owner.

- **Model type** (*categorical, between-subjects*). Each participant was randomly assigned to one of two configurations:

(1) Probabilistic. The AI system would calculate the probability of the reported address to be an illegal holiday rental based on a set of parameters. Each parameter was followed by a different number of (+) signs to indicate that some of those parameters had a more prominent impact on the final probability [15, 30].

(2) Rule-based. The AI system would evaluate whether the reported address meets relevant conditions that might indicate the property is being illegally rented as a holiday rental.

The parameters that the probabilistic and rule-based models would consider depend on the type of data that the AI system would retrieve. If publicly available data was retrieved, we would present participants with a few of the parameters that the original system suggested by the municipality of Amsterdam relies on for calculating a probability. If data that is not publicly available was retrieved, we would present participants with parameters related to the flow of people accessing the building.

- **Data provenance** (*categorical, between-subjects*). Each participant was randomly assigned to one of two configurations:

(1) Publicly-available data sources. The AI system would have access to and retrieve information available in the public registry.

(2) Non-publicly-available data sources. The AI system would have access to and retrieve the camera footage from the doorbell in the building or the footage from the nearest street camera.

5.1.2 Dependent Variables. The measurement instruments can be found in our repository.

- **Perceived ability**[12] (*continuous*). Measured by the average score on the six items suggested by Höddinghaus et al. [49].
- **Perceived benevolence** (*continuous*). Measured by the average score on the five items suggested by Mayer and Davis [69].
- **Perceived integrity** (*continuous*). Measured by the average score on the six items suggested by Mayer and Davis [69].
- **Perceived fairness** (*continuous*). Measured by a one-item construct on a 7-point Likert scale, following previous work [58, 60, 105].

5.1.3 Descriptive and Control Variables. The measurement instruments can be found in our repository.

- **Age group** (*categorical*). Age group that participants belong to. Participants chose one of the six categorical options.
- **Level of education** (*categorical*). Highest level of education that participants had completed. Participants chose one of the six categorical options.
- **Lessee of short-term rentals** (*categorical*). Experience renting out their property as a short-term rental —see Table 3.
- **AI literacy** (*continuous*). Knowledge and expertise working or interacting with AI [93]. We captured it through the average score on the four items suggested by Schoeffer et al. [93].
- **Affinity for technology** (*continuous*). Curiosity towards and willingness to engage with the technical working of systems [58]. We captured it through the average score on the four items suggested by Franke et al. [39], following previous work [58, 105].
- **Personal experience with decision-makers of illegal short-term rentals** (*continuous*). We captured participants' personal experience with algorithmic systems or humans making decisions about illegal holiday rentals through an adapted version of the

[11] The study was pilot-tested with 12 experts in human-computer interaction from our institution. During the pilot test, we checked the effectiveness of the manipulations, the feasibility of the presented scenarios [11], the layout, wording and potential biases that we might trigger [31].

[12] To validate if the responses of our participants were consistent with the initial definition and use of the measurement tools (i.e., items capturing perceived ability, benevolence, integrity) by Höddinghaus et al. [49] and Mayer and Davis [69], we conducted a principal component analysis (PCA). We encourage the interested reader to check the document *ABI-Fairness.pdf* (pages 46-49) in our repository.

Table 3: Overview of control variables and rationales for including them.

Control variable	Rationale for inclusion
Lessee of short-term rentals	We sought to understand whether having experience as a lessee of short-term rentals and, therefore, having a personal stake in the topic [21], had an impact on perceptions towards decision-makers.
AI literacy	It has been shown to impact fairness perceptions in algorithmic decision-making [93, 105].
Affinity for technology	It has been shown to affect perceptions of ability towards algorithmic systems [58].
Personal experience with decision-makers of illegal short-term rentals	Experience and familiarity with a specific decision-maker profile (algorithmic or non algorithmic) has been shown to lead to preferences towards that decision-maker [57].
Personal experience with public administration	From our qualitative study, we observed that, in 6 out of 21 participants, previous experiences with the public administration affected their perceptions towards the suggested scenarios.
Affinity for short-term rental policy	From our qualitative study, we observed that, in 4 out of 21 participants, perceptions towards the adequacy of the policy itself affected their perceptions towards the suggested scenarios.
Perceived task complexity	Previous work has shown that task complexity affects preferences towards human or algorithmic decision-makers [60, 77].

scale by Kramer et al. [57] and measured by the average score of the two suggested items.

- **Personal experience with public administration** (*continuous*). We employed an adapted version of the scale by Kramer et al. [57] and measured the average score on the two suggested items.
- **Affinity for short-term rental policy** (*continuous*). We measured affinity to policy through a one-item construct on a 7-point Likert scale, following previous work [74].
- **Perceived task complexity** (*continuous*). We measured perceived task complexity through a one-item construct on a 7-point Likert scale, similar to previous work [66, 105].

5.2 Procedure

We designed a four-step study —see Figure 5.

Step 1. Participants accepted the informed consent and responded to questions related to our exploratory variables (see section 5.1.3).

Step 2. Participants were shown a brief paragraph with information about the policy of their municipality in matters of short-term rentals. Participants were then introduced to the decision of the municipality to introduce an Artificial Intelligence system to accelerate the detection of illegal holiday rentals. Depending on which of the $2 \times 2 \times 2 = 8$ between-subject scenarios participants got randomly assigned to, they would read about a workflow where a fully-automated or a hybrid decision-maker configuration was put in place. Participants would also get to know whether the system relied on a probabilistic or rule-based model and whether it operated on publicly-available or non-publicly-available data. Participants would then be shown a graphical representation of the workflow[13] to facilitate comprehension.

Step 3. Participants were then shown an example of how the workflow looks in practice. The decision to do so was based on the observations from our qualitative study, where participants, especially those with lower AI literacy levels, would not understand what the jargon would entail in practice until they saw an example. Participants then answered the first attention check.

Step 4. Participants were asked to evaluate perceived ability, benevolence, and integrity towards the decision-maker through a set of questions (see section 5.1). After each set of questions, participants were asked to further elaborate and justify their perceptions of ability, benevolence, and integrity through open-ended questions.[14] The second attention check was located between the questionnaire about perceived ability and perceived benevolence. Participants were finally asked to evaluate their fairness perceptions towards the algorithmic decision-making process.

5.3 Data Collection

We planned to recruit at least 205 participants for data collection purposes. We calculated our planned sample by using the software *G*Power* [37], for a between-subjects ANOVA (*Fixed effects, special, main effects and interactions*). We calculated the sample size by setting the default effect size 0.25, a significance threshold of $\alpha = 0.05/7 = 0.007$ since we will test several hypotheses on the same data, a desired power of 0.8, with 8 groups and the respective degrees of freedom.

We recruited 223 participants on *Prolific* (https://www.prolific.com/) where we shared the link to our study with them. The study was conducted on *Qualtrics* (https://www.qualtrics.com/). All our participants were at least 18 years old and participated in the study only once. Since geographical location has been found to have an effect on fairness perceptions in algorithmic decision-making [52], we screened participants to ensure that they were located in a country in the Global North. All our participants were proficient in English. The participation in the study was compensated with an hourly rate of $12 or equivalent in the currency of the platform, which is higher than the federal minimum ($7.25/hour) and than the average compensation ($11/hour). Participants were introduced to an informed consent statement before they began the survey.

[13] The graphical representations for each scenario were designed so that participants would not anthropomorphize the algorithmic system or link human-like intelligence traits to it (e.g., by avoiding to represent the AI through a brain and a human-looking robot), as suggested by experts in the pilot study.

[14] For the sake of conciseness, we do not include the responses to open-ended questions in the main body of this paper. The interested reader can find these responses in our repository.

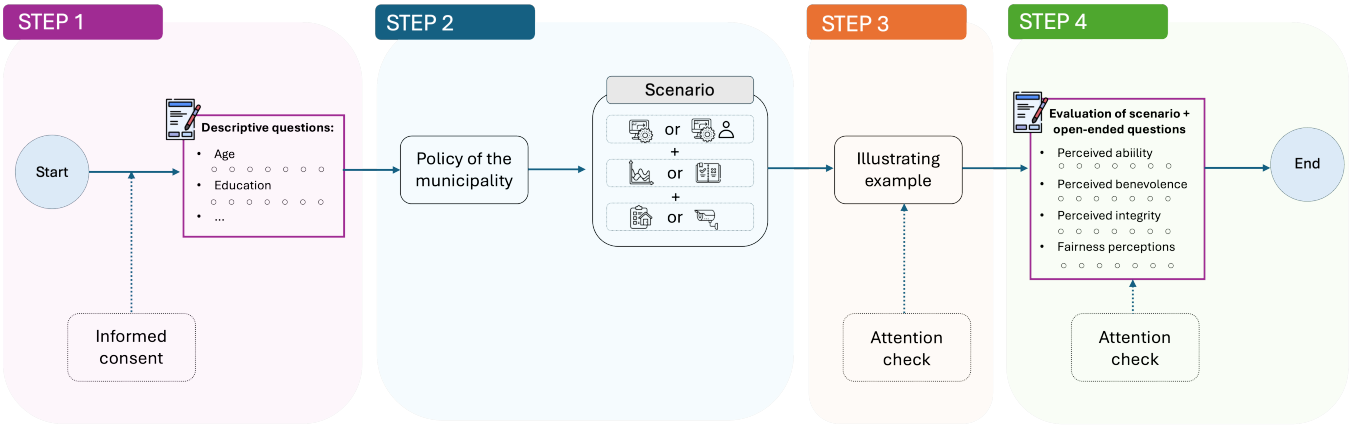


Figure 5: Procedure of the study.

5.4 Data Analysis

We mapped all (seven-point) Likert scale answers onto an ordinal scale going from −3 to 3 (i.e., from strongly disagree to strongly agree). We used both parametric and nonparametric tests in our analysis, and our choice of tests was informed by the criteria defined by Harwell [43]. We used parametric tests when the underlying assumptions of normality (Shapiro-Wilk test) and equality of variance (Bartlett's test) were satisfied, or when the test itself was robust to departures from these assumptions. For the sake of brevity, we will omit reporting the tests for assumptions. Since we are testing 7 hypotheses on the same data, we applied a Bonferroni correction to our significance threshold, reducing it to $\frac{0.05}{7} = 0.007$.

We used ANOVA (Analysis of Variance) as a parametric test and Kruskal-Wallis as a non-parametric test to examine the differences among the independent variables. Effect sizes for these tests were calculated using the eta-squared measure. We also used linear regression –both parametric and non-parametric– to model the influence of independent and control variables on dependent variables and to examine interaction effects. Finally, we conducted a mediation analysis to better explain the direct and indirect effects of the independent and control variables on the dependent variables. Mediation analysis [67] permits us to explore the nuanced effects of mediator variable(s) on the observed relationship between the independent (or control) and dependent variables – whether the observed *total* effect is the main effect or whether there is a *mediation* effect that can better explain the variance in the originally observed relationship.

6 Results and Analysis of Study 2

In this section we summarize the quantitative —confirmatory (section 6.2) and exploratory (section 6.3)— results of our study. The anonymized data, code for analysis (in R) and a report of the performed tests (with visualizations) are available in our repository.

6.1 Descriptive Statistics

For our study, we recruited 232 participants, out of which 223 participants passed both attention checks. Demographics are summarized in Table 4.

6.2 Hypothesis Tests

For our confirmatory analyses, we report the results for $\mathbf{H}_{1a}$, $\mathbf{H}_{1b}$, $\mathbf{H}_{1c}$, $\mathbf{H}_{1d}$ based on Kruskal-Wallis tests. To test **H2** we performed a non-parametric multiple linear regression.

$\mathbf{H}_{1a}$: We found a main effect of the decision-maker's profile on perceived ability, $\chi^2(1) = 72.01$, $p < .001$, $\eta^2 = 0.32$. Perceived ability was observed to be higher for hybrid profiles as compared to fully-automated ones (see Figure 6a).

$\mathbf{H}_{1b}$: We also found a main effect of the decision-maker's profile on perceived benevolence, $\chi^2(1) = 39.80$, $p < .001$, $\eta^2 = 0.18$. Perceived benevolence was found to be higher for hybrid profiles compared to fully-automated profiles (see Figure 6b). Even if the decision-maker's profile has a significant effect on perceived benevolence, it is worth noting that the mean values of perceived benevolence are below the midpoint of our chosen Likert scale of $[-3, +3]$; both for a hybrid decision-maker configuration (Mean = −0.49, Median = −0.6, SD = 1.49) and for a fully-automated one (Mean = −1.68, Median = −2.0, SD = 1.17).

$\mathbf{H}_{1c}$: We found no significant difference in perceived integrity across model type, $\chi^2(1) = 0.06$, $p > .1$, $\eta^2 = -0.004$.

$\mathbf{H}_{1d}$: We found no significant difference in perceived integrity based on the input data provenance, $\chi^2(1) = 2.69$, $p = .1$, $\eta^2 = 0.008$.

$\mathbf{H}_{2a}$, $\mathbf{H}_{2b}$, $\mathbf{H}_{2c}$: Our results showed that perceived abililty and integrity significantly affected fairness perceptions, however, the effect of perceived benevolence was not significant, $R^2 = 0.71$, $F(3, 219) = 93.35$, $\beta = 0.26$, $p < .001$. We observed that a unit increase in perceived ability resulted in a 0.42 point increase in fairness perceptions ($p < .001$). Similarly, a unit increase in perceived integrity led to a 0.63 point increase in fairness perceptions ($p < .001$).

We, therefore, found evidence in favor of four of our hypotheses ($\mathbf{H}_{1a}$, $\mathbf{H}_{1b}$, $\mathbf{H}_{2a}$, $\mathbf{H}_{2c}$). These results show that the decision-maker's profile has a main effect on both perceived ability and benevolence, and that perceived ability and perceived integrity relate positively to fairness perceptions.

6.3 Exploratory Analyses

Besides the pre-registered confirmatory analyses, we performed two types of analyses: (1) additional main and interaction effects of the

Table 4: Summary of our 223 participants' demographics.

Feature	Category (Number of participants, percentage)
Education	Incomplete high-school (1/223, 0.4%), High-school diploma (41/223, 18.4%), Some college education (52/223, 23.3%), Bachelor's degree (71/223, 31.8%), Professional Schooling (8/223, 3.6%), Postgraduate degree (50/223, 22.4%)
Age	19-25 years old (36/223, 16.14%), 26-35 years old (61/223, 27.36%), 28-50 years old (62/223, 27.8%), 50+ years old (74/223, 28.7%)
Self-reported AI literacy	Response to having a a good knowledge in the field of AI, working with AI, or being confident when interacting with AI: Disagreed (121/223, 54.26%), Agreed (102/223, 45.74%)

independent and control variables (see section 5.1.3) on perceived ability, benevolence, and integrity, and (2) mediation analyses as described earlier in section 5.4.

6.3.1 Main and Interaction Effects. We report the additional main and interaction effects we found through exploratory analyses.

Effect of Decision-Maker Profile on Perceived Integrity. Through a Kruskal-Wallis test, we examined differences in perceived integrity across profiles. Our analysis revealed that perceived integrity differed significantly across the decision-maker's profile, $\chi^2(1) = 53.08$, $p < .001$, $\eta^2 = 0.24$. Higher perceived integrity was reported for hybrid decision-maker configurations as compared to the fully-automated profile (see Figure 6a).

Effect of Model Type and Data Provenance on Integrity. In our confirmatory analyses, we found no significant main effect of model type or data provenance on perceived integrity. However, as an exploratory analysis, we examined the main and interaction effects of profile, model type, and data provenance on perceived integrity by fitting a linear regression. Our results indicate an interaction effect between model type and data provenance ($\beta = 0.83$, $p = .04$) in modeling perceived integrity, $R^2 = 0.26$, $F(7, 215) = 11.12$, $\beta = 0.27$, $p < .001$ (see Figure 7).

Effect of Policy Agreement on Perceived Integrity. Next, we examined the effect of participants' policy agreement on perceived integrity through a quantile regression. Our results showed a significant effect, $R^2 = 0.07$, $F(1, 221) = 20.07$, $\beta = 0.25$, $p < .001$. A one-point increase in policy agreement resulted in a 0.25-point increase in perceived integrity ($p = .03$). It is worth noting that although policy agreement has a significant effect on perceived integrity, the effect itself is weak.

6.3.2 Mediation Effects. We followed the procedure outlined by MacKinnon [67] in conducting the mediation analysis, and we tested the significance of the mediation effects using nonparametric bootstrapping approximations. Specifically, we computed unstandardized mediation effects for each of the 500 bootstrapped samples, and the 95% confidence interval (CI) was determined by computing the indirect effects at the 2.5^{th} and 97.5^{th} percentiles.

Mediation Effect of Perceived Ability on the Relationship Between Decision-Maker's Profile and Perceived Fairness. In section 6.2 we reported significant effects of profile on perceived ability, and of perceived ability on fairness perceptions. Consequently, we hypothesize that these two effects may be related and that perceived ability may mediate the effect of profile on fairness perceptions. We observed that the regression coefficients between profile and fairness perceptions ($\beta = 1.25$, $p < .001$), and between perceived ability and fairness perceptions ($\beta = 0.80$, $p < .001$) were significant (see Figure 8a) . In addition, we observed a complete and significant mediation effect, $\beta = 1.40$, $CI = [1.07, 1.77]$, $p < .001$.

Mediation Effect of Perceived Integrity on the Relationship Between Decision-Maker's Profile and Perceived Fairness. As with perceived ability, we hypothesize that perceived integrity may mediate the effect of profile on fairness perceptions. Our analysis revealed another complete and significant mediation effect, $\beta = 1.23$, $CI = [0.94, 1.51]$, $p < .001$. The regression coefficients between profile and fairness perceptions ($\beta = 1.25$, $p < .001$), and between perceived integrity and fairness perceptions ($\beta = 1.08$, $p < .001$) were significant (see Figure 8b).

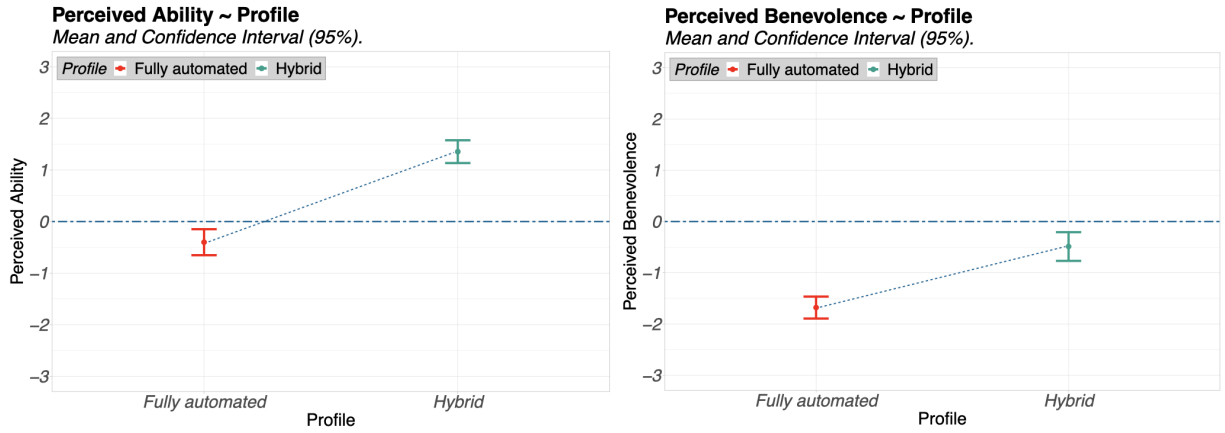
Mediation Effect of Perceived Integrity on the Relationship Between Policy Agreement and Perceived Fairness. Previously, we reported a significant effect of policy agreement on perceived integrity. In addition, our exploratory analysis revealed a significant effect of policy agreement on fairness perceptions, $R^2 = 0.02$, $F(1, 221) = 5.18$, $\beta = 0.45$, $p = .02$. Therefore, we conducted a mediation analysis with perceived integrity as the mediator. Our results show a significant mediation effect, $\beta = 0.20$, $CI = [0.05, 0.37]$, $p = .008$. The regression coefficients between policy agreement and fairness perceptions ($\beta = 0.18$, $p = .02$) and between perceived integrity and fairness perceptions ($\beta = 1.09$, $p < .001$) were significant (see Figure 9).

7 Discussion

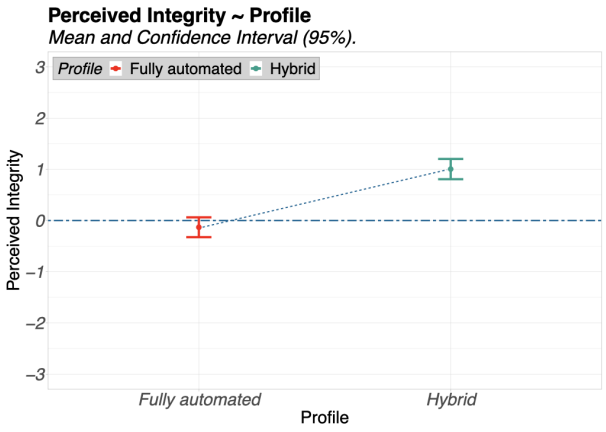
Drawing from our findings and prior literature, we discuss implications for the design of algorithmic decision-making processes in the public sector and for future HCI research.

7.1 Summary of Results In Relation to Previous Work

In this section, we summarize the results of our interview **(RQ1.1., RQ1.2.)** and large-scale quantitative studies **(RQ2.1., RQ2.2.)**. We focus on the findings related to the decision-maker profile in section 7.1.1, and on the findings related to the model type and data provenance in section 7.1.2.



(a) Effect of decision-maker's profile on perceived ability. **(b) Effect of decision-maker's profile on perceived benevolence.**



(c) Effect of decision-maker's profile on perceived integrity.

Figure 6: This figure illustrates the significant effects of decision-maker's profile on perceived ability, benevolence, and integrity.

7.1.1 Effect of decision-maker profile. In algorithmic decision-making, human intervention aims at ensuring that decisions are not uniquely based on decision-subjects' *data shadows*, i.e., computational representations of decision-subjects through aspects of a person that can be metrified [73]. Findings from our interviews indicate that decision-subjects' perspectives on human intervention were aligned with such intention, profile of decision-makers (i.e., with or without human intervention) being a prominent characteristic that decision-subjects would consider when assessing the adequacy of decision-maker configurations (**RQ1.1.**). Decision-subjects evaluated the ability and benevolence of decision-maker configurations based on the decision-maker's profile (**RQ1.2.**). In our quantitative *between-subjects* study, and unlike previous work [77, 103, 105], we *did* find statistically significant differences between ability and benevolence perceptions towards hybrid decision-maker configurations and fully-automated ones, hybrid configurations being perceived as more able and benevolent (**RQ2.1.**). Additionally, our results indicate that there might be an effect of decision-maker profile on integrity perceptions too, hybrid configurations being associated with higher levels of integrity. The reason why we found significant differences between decision-makers' profiles might be due to (1) presenting a hybrid decision-maker configuration where the interaction paradigm relies on advisory control rather than supervisory control [89] and (2) differences in research method. Previous work comparing decision-subjects' perceptions towards fully-automated *vs.* hybrid decision-maker configurations [77, 103, 105] mainly gave humans a supervisory role and attributed them the task to monitor AI's actions (supervisory control [89]). Instead, we explicitly indicated that the AI's task was limited to flagging potential illegal holiday rentals but it was the human who would evaluate the output and make the final decision (advisory control [89]). The advisory control paradigm might have led participants to perceiving human intervention as more effective. On a methodological level, we followed practices from literature in organizational psychology for human decision-making [25]. Instead of capturing the effect of decision-maker configurations (a) on fairness perceptions directly [96], or (b) through fairness scales with little emphasis on the decision-maker [24], we first measured decision-subjects'

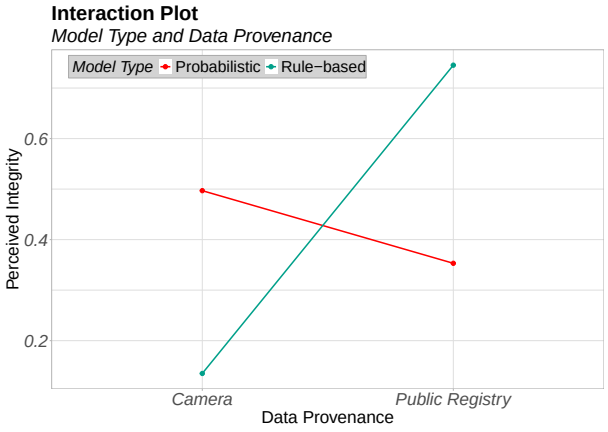


Figure 7: This figure shows the significant interaction effect between *model type* and *data provenance* when modeling *perceived integrity*.

perceptions of ability, benevolence, and integrity towards decision-makers. We then captured fairness perceptions towards algorithmic decision-making. Such an approach enabled us to verify that decision-maker configurations with human intervention were seen as more able and benevolent, and associated with higher levels of integrity than fully-automated configurations.

It should be noted that, even if hybrid decision-maker configurations were perceived as more benevolent than fully-automated ones, benevolence perceptions were still negative in every case we evaluated. We suspect the nature of the public sector might have had an impact on such results. Algorithmic decision-making in the public sector presents several peculiarities compared to the private sector [2]. Unlike the private sector, where decision-subjects can, e.g., look for an alternative financial company if their loan gets rejected [105], decision-subjects necessarily have to deal with decisions made by public institutions [2]. This lack of alternatives might have contributed to negative benevolence perceptions across conditions. Additionally, participants might have perceived that, even when a human was making the final evaluation informed by the AI, the suggested action (i.e., sending a warning) was too harsh.

7.1.2 Effect of Model Type and Data Provenance. In our interview study, we observed that model type and input data provenance were also prominent characteristics that decision-subjects would consider when assessing the adequacy of decision-maker configurations (**RQ1.1.**) Participants evaluated decision-makers' integrity based on the model type and the input data provenance (**RQ1.2.**). Interviewees were especially interested in receiving a clear statement about the *cause* that led to the warning and the rules that they, as decision-subjects, had violated (i.e., they were asking for a justification [45]). The lack of alternatives and the nature of the public sector might also explain this demand, which would align with findings by Aljuneidi et al. [2]. Aljuneidi et al. [2] observed requests for justifications in a scenario capturing decision-subjects' fairness perceptions towards an algorithmic process for expired ID-card renewals. Instead, for a loan approval scenario in the private sector, counterfactual explanations were considered adequate as long as these were actionable [93] —without necessarily having to point to the appropriateness of the factors, which is needed in *justifications* [45].

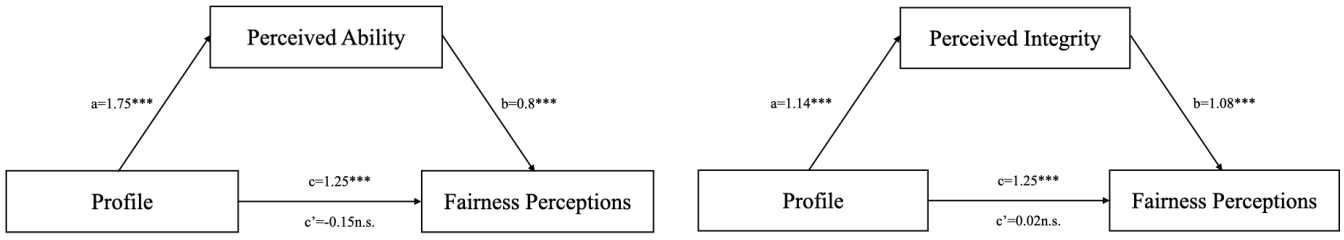
In our quantitative study, we did not find a main effect of model type and data provenance on perceived integrity (**RQ2.1.**). However, our quantitative study did reveal that there might be an interaction effect between model type and data provenance when predicting perceptions of integrity. The desire for systems that provide justifications like rule-based models, therefore, depends on the data source (publicly available or non-publicly available) that the model relies on. This suggests that, even for contexts such as policy enforcement, relying on data that respects decision-subjects' privacy is key in shaping decision-subjects' perceptions. Exploratory results also indicate that, in addition to model type and input data provenance, decision-subjects' agreement with the implemented policy might have an effect on integrity perceptions, which mediates its effect on fairness perceptions.

Findings from our large-scale quantitative study also showed that perceptions of ability and integrity relate positively to fairness perceptions **(RQ2.2.)**. We further discuss this finding in section 7.4.

7.2 Implications for Designing Algorithmic Decision-Making Processes in the Public Sector

Based on our findings, we highlight four main recommendations for designers developing and deploying AI for public decision-making.

(1) **Design workflows where street-level bureaucrats can meaningfully intervene in algorithmic decision-making.** Our study suggests that, when humans are meaningfully involved in the algorithmic decision-making process, decision-subjects' perceptions of ability, benevolence, and integrity towards the decision-maker tend to improve. Therefore, the first design implication is that decision-making workflows should be structured to ensure street-level bureaucrats are actively involved and maintain effective control when interacting with AI (e.g., through an advisory control paradigm [89]). However, even if street-level bureaucrats have final control over decisions, the nature of their interaction with the AI requires careful consideration. Prior work has highlighted the problematic opacity of AI systems [19, 88], i.e., presenting high-dimensional characteristics stemming from mathematical optimizations in a format adapted to end-users' needs for semantic interpretation and reasoning is not a trivial task [19]. Difficulties in presenting algorithmic outputs might lead to overreliance on the AI system [17, 51]. End-users' cognitive biases have also been shown to contribute to overreliance [44]. Public agencies designing AI systems and integrating them in decision-making should, therefore, carefully look into how interactions between street-level bureaucrats and AI systems occur so that street-level bureaucrats can apply their tacit knowledge when making decisions [4]. This is necessary to prevent human intervention from boiling down to a confirmation mechanism of algorithmic outputs [91]. Explanations [102], cognitive forcing functions [17], or reinforcement learning paradigms [18, 53] have been suggested as potential solutions to AI overreliance. For algorithmic decision-making in the public sector, street-level bureaucrats



(a) Mediation effect of perceived ability on the relationship between decision-maker's profile and fairness perceptions.

(b) Mediation effect of perceived integrity on the relationship between the decision-maker's profile and fairness perceptions.

Figure 8: This figure illustrates the mediation effects of perceived ability and integrity on the relationships between the decision-maker's profile and their fairness perceptions.

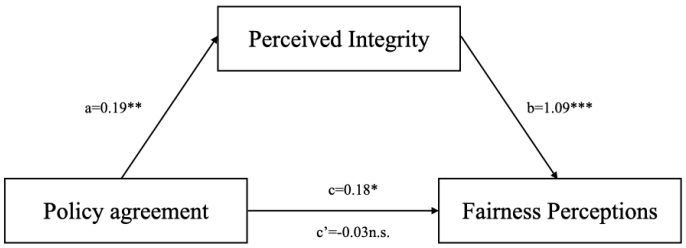


Figure 9: This figure shows the mediation effect of *perceived integrity* on the relationship between policy agreement and fairness perceptions.

might be better positioned to apply discretion if they were provided with multidimensional outputs and algorithmic *suggestions* instead of mandated outcomes [92]. Designers should evaluate the utility of those solutions while considering the complex bureaucratic processes street-level bureaucrats face in their everyday practice [107].

(2) **Balance the need for justifications and decision-subjects' right to privacy.** Our quantitative results showed an interaction effect between data provenance and model type. This indicates that decision-subjects' wish for justifications (which would indicate their preference toward rule-based models) does not hold when compliance with existing rules is evaluated based on data that comes from ethically questionable sources. Public organizations designing future algorithmic decision-making processes for policy enforcement should, therefore, balance the need to rely on models that provide justifications about the decision and the need to respect decision-subjects' privacy, i.e., rule-based AI systems should not be implemented when the data that these systems evaluate does not align with the principle of proportionality.

(3) **Disentangle perceptions towards hybrid decision-maker configurations and perceptions towards the implemented policy.** Our exploratory quantitative findings indicate that integrity perceptions towards decision-maker configurations might be impacted by participants' *agreement with the policy* behind the identification of illegal short-term rentals. This finding implies that public institutions aiming to inform effective mechanisms for human intervention by capturing decision-subjects' fairness perceptions should disentangle decision-subjects' perceptions towards the suggested mechanisms and their agreement with the enforced policies. This requires crafting experimental designs that not only capture perceptions towards human-AI configuration properties, but also towards the alignment between the goal of the decision-making and citizens' political stance. Representative modes of civic participation are well suited to ensure that the enforced policies are aligned with democratic values [1].

(4) **Engage with impacted communities when designing human intervention in algorithmic decision-making processes.** Beyond a mere quality control mechanism, human intervention should represent an *effective* means for protecting decision-subjects' fundamental rights (e.g., human dignity) [4]. It is, therefore, important that organizations developing and deploying AI systems for public decision-making account for the perceptions towards human intervention of communities who will suffer the consequences of automating those processes [48]. Ours is an effort in this direction. Recent studies indicate that cities like Amsterdam include civic participation approaches to inform the design of pilot AI systems [1]. If municipalities like Amsterdam were to integrate our approach as part of their civic participation initiatives, we recommend that they engage with individuals who have previously been impacted by similar systems or, who might be impacted in the future in that specific municipality. Through interviews, designers could capture impacted communities' lived experiences, which would help

identify additional factors that contribute to perceptions of fairness for that specific context. There might be cultural factors that our study has not captured and that are relevant for that case. The qualitative insights could then be complemented with a large-scale quantitative user study for capturing perceptions of citizens of that municipality. This would shed light on the generalizability of the qualitative findings and on the broader acceptance of the suggested decision-maker configuration. Studies like these would address the need to encourage public participation and reasoned deliberation about public AI, moving away from procurement processes with limited visibility of design choices [76].

HCI scholars could additionally contribute in this direction by examining how human intervention is being shaped in real-world public algorithmic decision-making processes. This includes exploring (a) whether and how participatory approaches focus on informing human intervention, (b) what mechanisms exist for scaffolding decision-subjects' perceptions when shaping human intervention, or (c) how to adapt existing (generic) frameworks for responsible AI design to specifically focus on human intervention design [27, 33]. Exploring how human intervention is shaped is especially relevant in an era where the European Union's Artificial Intelligence Act (entered into force on August 1st 2024) will require deployers of high-risk AI systems to provide a "description of the implementation of human oversight measures" as part of a "Fundamental Rights Impact Assessment" (Article 27(1)) [36].

7.3 Making Complex and Distributed Human Intervention(s) Visible Across AI Pipelines

Our findings confirm the need for street-level bureaucrats to retain discretionary power to effectively intervene in algorithmic decision-making and to safeguard decision-subjects' rights. However, those designing decision-support AI systems *also* hold some level of discretionary power [107]. By translating high-level system goals into specific design requirements, system designers encode legislation into software [107]. Findings from our interviews indicate that most of our participants thought of human intervention as the act of providing human input *at the time of decision-making* for correcting AI errors (aligned with how the GDPR [35] defines human intervention). Only a few were interested in knowing how humans intervene in the early stages of AI design. It could be argued, however, that human intervention in algorithmic decision-making should not be limited to the human making the final decision. Instead, human intervention should account for the complex and distributed human labor that AI systems result from [91], i.e., human intervention should be framed as a *problem of many hands* [23]. This requires to acknowledge the (partial) shift of discretionary power from decision-making time to design time [107], and to ensure *reflexivity* at all stages of the AI pipeline. The HCI community could explore several future research directions stemming from a holistic take on human intervention.

One of those future research directions involves adopting a *preventive approach to human intervention* [4]. There are different ways of "datafying" an action or a person [47]. A preventive approach to human intervention [4] advocates for disclosing and challenging the assumptions underneath design choices (and the rationales that led to those choices [104]). Practitioners need both (1) infrastructure [7] and (2) guidance [9, 28, 68] to meaningfully exercise reflexivity. Future HCI research could look into methods for bringing visibility to the design choices (e.g., choices on which data to include or not to include when training AI systems [75]) that shape machine behaviour [83, 84, 90] and the downstream impact of such choices.

Furthermore, with the proliferation of generative AI systems, AI design pipelines are becoming increasingly modular [8, 23]. Since actors distributed across different organizations contribute to the production, deployment and use of AI systems, responsibility is distributed across those actors and there is limited visibility of the choices made by others (i.e., actors suffer from *accountability horizon* [23]). Future HCI research should further investigate the dynamics that prevail in those algorithmic supply chains. This includes conducting ethnographic and workplace studies to uncover, e.g., who is involved in algorithmic supply chains, how their interactions are structured, or how AI supply chains develop over time [8, 23].

7.4 Adapting the ABI Model to Algorithmic Decision-Making

To capture decision-subjects' perceptions towards algorithmic decision-maker configurations with varying levels of human intervention, we characterized each decision-maker configuration based on Mayer et al. [70]'s ability, benevolence, and integrity (ABI) model. We then related perceptions of ability, benevolence, and integrity to decision-subjects' fairness perceptions. Our confirmatory analysis showed that perceived ability and integrity positively relate to fairness perceptions. Our exploratory analyses further revealed a mediation of both perceived ability and integrity on the effect that decision-makers' profile has on fairness perceptions. Similarly, a mediation analysis revealed that the effect of policy agreement on fairness perceptions might be *mediated* by perceived integrity. These results are testimony to the potential suitability of the multidimensional ABI model [70] to provide a nuanced understanding of how and why fairness perceptions towards algorithmic decision-making processes might be mediated by decision-subjects' perceptions towards decision-makers.

The ABI model [70] was created to capture perceived trustworthiness (conceptualized through perceptions of ability, benevolence, and integrity) towards *human* decision-makers [70]. Even if not explicitly developed for algorithmic decision-making, using the ABI model [70] was especially suitable in our study because it distinguishes perceptions towards decision-makers from trustor-related and contextual factors. This brings conceptual clarity and precision when capturing the relationship between perceptions toward decision-makers and fairness perceptions in algorithmic decision-making. We followed Höddinghaus et al. [49] and modified the dimension of *ability* to highlight data processing capabilities and flexibility in algorithmic decision-making. The dimensions of *benevolence* and *integrity* were captured through the tool developed by

Mayer and Davis [69]. In light of our findings, future research capturing decision-subjects' perceptions towards algorithmic decision-maker configurations could benefit from adopting an approach similar to ours. However, further methodological contributions are needed to capture the unique parameters that define benevolence and integrity in algorithmic decision-making. Although efforts in this direction have taken place from an *end user* perspective in the area of automation [14, 42, 54, 56, 59, 72, 94] (see section 2.3), the need for adapting the ABI model [70] from the perspective of *decision-subjects* or the *wider public* has received relatively little attention. From the interviews, for example, we identified that, for algorithmic decision-making, explainability and actionability of the decision basis could be important parameters within the dimension of integrity (see section 3.2). Methodological approaches are needed to systematically identify factors unique to algorithmic decision-making and rigorously validate constructs equivalent to the ABI model [70] across different contexts.

7.5 Caveats and Limitations

In this section, we discuss relevant caveats and report the limitations of our study.

(1) *Participants With a Personal Stake:* For our qualitative study, we decided to recruit participants with experience renting their properties out as short-term rentals. We did so to ensure our participants had a personal stake in the hypothetical scenario [21]. For our main study, instead, we did not screen participants based on their experience as short-term rental lessors. We decided to tell the story in the third person, asked participants to look into the scenario through the lens of a decision-subject [93, 105], and captured participants' experience renting properties out as short-term rentals as a control variable (section 5.1.3). We did so to avoid *outcome favorability bias* [65, 103] as it has been done in prior work [6, 93, 95]. We suspect results might vary if all participants had experience with short-term rentals, e.g., the perceived of appropriateness of the enforced policy might be lower, affecting integrity perceptions.
(2) *Participants With Different Cultural Backgrounds:* We recruited participants from the Global North who were proficient in English. Fairness perceptions towards algorithmic decision-making have been shown to vary depending on whether participants belong to the Global North or South [52]. Our study might, therefore, be subject to representativeness limitations [61].
(3) *Additional Characteristics and Human Factors:* Our study controlled for a limited number of decision-maker characteristics and human factors. However, additional characteristics (e.g., training data) or human factors (e.g., AI skepticism) might impact decision-subjects' perceptions towards algorithmic decision-maker configurations in different cultural contexts and use cases.
(4) *Generalizability Across Use Cases:* Our study is limited to a single use case (i.e., detection of illegal holiday rentals) to generate in-depth insights into the selected context [21]. We expect our results to partially generalize to other use cases. We expect the effect of hybrid decision-maker configurations on perceptions of ability, benevolence and integrity to be generalizable across use cases as long as the presented human intervention is as meaningful as in an advisory control paradigm [89]. We expect negative benevolence perceptions and the interaction between data provenance and model type to generalize only to other policy enforcement contexts. For contexts other than policy enforcement, however, statistical inferences that provide counterfactual explanations may be perceived as acceptable [93] and lead to positive integrity perceptions regardless of the data provenance. As for the effect of policy agreement on perceptions of integrity, we expect this effect to be generalizable to use cases beyond policy enforcement. While in contexts other than policy enforcement there is no "implemented policy" as such, we predict that the agreement with the political principles inherent to a specific decision-making process may affect perceived integrity. For example, in a loan approval process, decision-subjects' perception towards the need to request a loan in itself –instead of the government offering every citizen a home– may affect perceptions of integrity towards the decision-maker configuration.
(5) *Effect of Design Choices:* We made specific design choices when selecting the terminology and designing the visual stimuli for our quantitative study. We decided to use the term Artificial Intelligence and avoid images that anthropomorphize algorithmic systems (e.g., brains, humanoid robots). Results might have been different if we had used a different terminology (e.g., computational system, statistical model) [58] or visual means.

8 Conclusion

Human intervention aims at safeguarding decision-subjects' "rights, freedoms, and legitimate interests" [35] in algorithmic decision-making. While hybrid decision-maker configurations (i.e., algorithmic decision-making with human intervention) are claimed to combine the efficiency and data processing capabilities of AI systems with the flexibility of humans, there is little empirical evidence showing that decision-subjects perceive these as fairer than fully-automated decision-maker configurations. This paper presented a mixed-method study to evaluate whether human intervention effectively contributes to decision-subjects' fairness perceptions in an algorithmic illegal holiday rental detection scenario. Through a foundational interview study, we first identified *decision-maker profile*, *model type*, and *input data provenance* as three influential characteristics that might co-shape decision-subjects' perceptions towards decision-maker configurations. Through a large-scale quantitative study, we then tested the effect of those three characteristics on *perceived ability, benevolence*, *integrity* (ABI) and on *fairness* perceptions. We found that *decision-maker's profile* affect perceived ability and benevolence, and might also affect integrity. Our results also showed that the ABI model [70] might be a useful instrument to capture perceptions towards decision-maker configurations: perceived ability and integrity not only contribute positively to fairness perceptions, but they might mediate the effect of decision-maker's profile on fairness perceptions.

Our results provide empirical evidence that human intervention is, indeed, effective in improving decision-subjects' perceptions towards algorithmic decision-maker configurations. Based on our findings, we suggest four main recommendations for designers developing and deploying public AI systems for decision-making.

Our findings additionally encourage the HCI community to inspect current real-world practices in shaping human intervention across algorithmic supply chains. Our work, therefore, informs the design of *effective* human intervention.

Acknowledgments

We thank the participants in our interview and crowdsourced studies. We also thank our colleagues at the Knowledge and Intelligence Design group for helping us pilot test our study.

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990.

References

[1] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. (2 2023). doi:10.1145/3544548.3580984

[2] Saja Aljuneidi, Wilko Heuten, Larbi Abdenebaoui, Maria K Wolters, and Susanne Boll. 2024. Why the Fine, AI? The Effect of Explanation Level on Citizens' Fairness Perception of AI-based Discretion in Public Administrations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 318, 18 pages. doi:10.1145/3613904.3642535

[3] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300760

[4] Marco Almada. 2019. Human intervention in automated decision-making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ACM, New York, NY, USA, 2–11. doi:10.1145/3322640.3326699

[5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3411764.3445736

[6] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (9 2020), 611–623. doi:10.1007/s00146-019-00931-w

[7] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal* 30, 5 (9 2021), 739–768. doi:10.1007/s00778-021-00671-8

[8] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2024. An Empirical Exploration of Trust Dynamics in LLM Supply Chains. *arXiv preprint arXiv:2405.16310* (2024).

[9] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 482–495. doi:10.1145/3600211.3604674

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2020. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. (4 2020).

[11] Jeffrey Bardzell, Shaowen Bardzell, and Erik Stolterman. 2014. Reading critical designs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1951–1960. doi:10.1145/2556288.2557137

[12] Patrick Bedué and Albrecht Fritzsche. 2022. Can we trust AI? an empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management* 35, 2 (2022), 530–549.

[13] Gianluca Bei and Filippo Celata. 2023. Challenges and effects of short-term rentals regulation. *Annals of Tourism Research* 101 (7 2023), 103605. doi:10.1016/j.annals.2023.103605

[14] Izak Benbasat and Weiquan Wang. 2005. Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems* 6, 3 (3 2005), 72–101. doi:10.17705/1jais.00065

[15] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. (1 2018). doi:10.1145/3173574.3173951

[16] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3290605.3300271

[17] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (4 2021), 1–21. doi:10.1145/3449287

[18] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Susan A Murphy, and Krzysztof Z Gajos. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. *arXiv preprint arXiv:2403.05911* (2024).

[19] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016).

[20] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (10 2019), 809–825. doi:10.1177/0022243719851788

[21] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners*. Sage publications ltd. 1–400 pages.

[22] Victoria Clarke and Virginia Braun. 2021. *Thematic analysis: a practical guide*. SAGE Publications Ltd.

[23] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1186–1197. doi:10.1145/3593013.3594073

[24] Jason A. Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86, 3 (6 2001), 386–400. doi:10.1037/0021-9010.86.3.386

[25] Jason A. Colquitt and Jessica B. Rodell. 2011. Justice, Trust, and Trustworthiness: A Longitudinal Analysis Integrating Three Theoretical Perspectives. *Academy of Management Journal* 54, 6 (12 2011), 1183–1206. doi:10.5465/amj.2007.0572

[26] John Cook and Toby Wall. 1980. New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of occupational psychology* 53 (1980), 39–52. Issue 1.

[27] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. doi:10.1145/3544548.3581026

[28] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 473–484. doi:10.1145/3531146.3533113

[29] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. doi:10.1037/xge0000033

[30] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. (1 2019). doi:10.1145/3301275.3302310

[31] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (10 2021), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

[32] Tim Draws, Zoltán Szlávik, Benjamin Timmermans, Nava Tintarev, Kush R. Varshney, and Michael Hind. 2021. Disparate Impact Diminishes Consumer Trust Even for Advantaged Users. (1 2021). doi:10.1007/978-3-030-79460-6{_}11

[33] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.

[34] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA.

[35] European Commission. 2018. 2018 reform of EU data protection rules. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

[36] European Commission. 2021. Regulation of the European Parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[37] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (5 2007), 175–91. doi:10.3758/bf03193146

[38] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (5 2019), 261–262. doi:10.1038/s42256-019-0055-y

[39] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer*

Interaction 35, 6 (4 2019), 456–467. doi:10.1080/10447318.2018.1456150

[40] Elena Fumagalli, Sarah Rezaei, and Anna Salomons. 2022. OK computer: Worker perceptions of algorithmic recruitment. *Research Policy* 51, 2 (3 2022), 104420. doi:10.1016/j.respol.2021.104420

[41] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS SYMPOSIUM ON MACHINE LEARNING AND THE LAW 8*.

[42] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38, 10 (10 2019), 1004–1015. doi:10.1080/0144929X.2019.1656779

[43] Michael R Harwell. 1988. Choosing between parametric and nonparametric tests. *Journal of Counseling & Development* 67, 1 (1988), 35–38.

[44] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. doi:10.1145/3544548.3581025

[45] Clément Henin and Daniel Le Métayer. 2021. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* (7 2021). doi:10.1007/s00146-021-01251-8

[46] César Hidalgo, Diana Orghian, Jordi Albo-Canals, Filipa de Almeida, and Natalia Martin. 2021. *How Humans Judge Machines*. MIT Press. https://hal.archives-ouvertes.fr/hal-03058652

[47] Mireille Hildebrandt. 2017. Profiles and correlatable humans. In *Who Owns Knowledge?* Routledge, 265–284.

[48] Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law* 20, 1 (2019), 83–121.

[49] Miriam Höddinghaus, Dominik Sondern, and Guido Hertel. 2021. The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior* 116 (3 2021), 106635. doi:10.1016/j.chb.2020.106635

[50] Nataliya V Ivankova and John W Creswell. 2009. Mixed methods. *Qualitative research in applied linguistics: A practical introduction* 23 (2009), 135–161.

[51] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11 (2 2021), 108. Issue 1. doi:10.1038/s41398-021-01224-x

[52] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. doi:10.1145/3491102.3517533

[53] Daniel N Kluttz, Nitin Kohli, and Deirdre K Mulligan. 2022. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics*. Auerbach Publications, 420–428.

[54] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12 (10 2021). doi:10.3389/fpsyg.2021.604977

[55] Daan Kolkman. 2020. The usefulness of algorithmic models in policy making. *Government Information Quarterly* 37, 3 (7 2020), 101488. doi:10.1016/j.giq.2020.101488

[56] Sherrie Yi Xiao Komiak. 2003. *The impact of internalization and familiarity on trust and adoption of recommendation agents*. Ph. D. Dissertation. https://open.library.ubc.ca/collections/831/items/1.0091325

[57] Max F. Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. 2018. When Do People Want AI to Make Decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 204–209. doi:10.1145/3278721.3278752

[58] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–28. doi:10.1145/3491102.3517527

[59] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (1 2004). doi:10.1518/hfes.46.1.50{_}30392

[60] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (1 2018). doi:10.1177/2053951718756684

[61] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. doi:10.1145/3411764.3445570

[62] Roy J Lewicki and Barbara Benedict Bunker. 1995. *Trust in relationships: A model of development and decline*. Jossey-Bass/Wiley.

[63] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (3 2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[64] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46, 4 (12 2019), 629–650. doi:10.1093/jcr/ucz013

[65] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 764–774. doi:10.1145/3593013.3594041

[66] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. doi:10.1145/3491102.3517606

[67] David MacKinnon. 2012. *Introduction to statistical mediation analysis*. Routledge.

[68] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (3 2022), 1–26. doi:10.1145/3512899

[69] Roger C. Mayer and James H. Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology* 84, 1 (2 1999), 123–136. doi:10.1037/0021-9010.84.1.123

[70] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (7 1995), 709. doi:10.2307/258792

[71] D. J. McAllister. 1995. AFFECT- AND COGNITION-BASED TRUST AS FOUNDATIONS FOR INTERPERSONAL COOPERATION IN ORGANIZATIONS. *Academy of Management Journal* 38, 1 (2 1995), 24–59. doi:10.2307/256727

[72] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (9 2002), 334–359. doi:10.1287/isre.13.3.334.81

[73] Isak Mendoza and Lee A Bygrave. 2017. The right not to be subject to automated decisions based on profiling. *EU internet law: Regulation and enforcement* (2017), 77–98.

[74] Ryan Merrill and Nicole Sintov. 2016. An Affinity-to-Commons Model of Public Support For Environmental Energy Policy. *Energy Policy* 99 (12 2016), 88–99. doi:10.1016/j.enpol.2016.09.048

[75] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–19. doi:10.1145/3491102.3517644

[76] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.

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

[78] David T. Newman, Nathanael J. Fast, and Derek J. Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (9 2020), 149–167. doi:10.1016/j.obhdp.2020.03.008

[79] Shirley Nieuwland and Rianne van Melik. 2020. Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals. *Current Issues in Tourism* 23, 7 (4 2020), 811–825. doi:10.1080/13683500.2018.1504899

[80] Ronald C. Nyhan and Herbert A. Marlowe. 1997. Development and Psychometric Properties of the Organizational Trust Inventory. *Evaluation Review* 21 (10 1997), 614–635. Issue 5. doi:10.1177/0193841X9702100505

[81] Gideon Ogunniye, Benedicte Legastelois, Michael Rovatsos, Liz Dowthwaite, Virginia Portillo, Elvira Perez Vallejos, Jun Zhao, and Marina Jirotka. 2021. Understanding User Perceptions of Trustworthiness in E-Recruitment Systems. *IEEE Internet Computing* 25, 6 (11 2021), 23–32. doi:10.1109/MIC.2021.3115670

[82] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (3 2022), 1–31. doi:10.1145/3512929

[83] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018), 1–28. doi:10.1145/3274405

[84] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (7 2020), 205395172093960. doi:10.1177/2053951720939605

[85] Jeanne S Ringel, Dana Schultz, Joshua Mendelsohn, Stephanie Brooks Holliday, Katharine Sieck, Ifeanyi Edochie, and Lauren Davis. 2018. Improving child welfare outcomes: balancing investments in prevention and treatment. *Rand health quarterly* 7, 4 (2018).

[86] Karlene H Roberts and Charles A O'Reilly. 1974. Measuring organizational communication. *Journal of applied psychology* 59 (1974), 321. Issue 3.

[87] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review* 23 (1998), 393–404. Issue 3. http://www.jstor.org/stable/259285

[88] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (5 2019), 206–215. doi:10.1038/s42256-019-0048-x

[89] Shadan Sadeghian, Alarith Uhde, and Marc Hassenzahl. 2024. The Soul of Work: Evaluation of Job Meaningfulness and Accountability in Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 130 (April 2024), 26 pages. doi:10.1145/3637407

[90] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. doi:10.1145/3411764.3445518

[91] Claudio Sarra et al. 2020. Defenceless? An analytical inquiry into the right to contest fully automated decisions in the GDPR. *An Anthology of Law* (2020), 235–252.

[92] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–41. doi:10.1145/3476089

[93] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. (5 2022). doi:10.1145/3531146.3533218

[94] Elizabeth Solberg, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. 2022. A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids. *Group & Organization Management* 47, 2 (4 2022), 187–222. doi:10.1177/10596011221081238

[95] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA, 2459–2468. doi:10.1145/3292500.3330664

[96] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (7 2022), 205395172211151. doi:10.1177/20539517221115189

[97] Tom R Tyler. 1989. The psychology of procedural justice: A test of the group-value model. *Journal of personality and social psychology* 57, 5 (1989), 830.

[98] Tom R. Tyler and E. Allan Lind. 1992. A Relational Model of Authority in Groups. 115–191. doi:10.1016/S0065-2601(08)60283-X

[99] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants". *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020), 1–22. doi:10.1145/3415238

[100] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3411764.3445365

[101] Niels van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170 (2 2023), 102954. doi:10.1016/j.ijhcs.2022.102954

[102] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (4 2023), 1–38. doi:10.1145/3579605

[103] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. doi:10.1145/3313831.3376813

[104] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities. In *2023 Human-Centered XAI Workshop at CHI Conference on Human Factors in Computing Systems (CHI '23)*.

[105] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. doi:10.1145/3544548.3581161

[106] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–28. doi:10.1145/3491102.3517791

[107] Stavros Zouridis, Marlies van Eck, and Mark Bovens. 2020. *Automated Discretion*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-19566-3_20