Understanding Decision Subjects' Needs and Perceptions Towards Contestable AI Systems

Yurrita Semperena, M.

DOI
10.4233/uuid:5792abf8-c8ff-4301-ba06-e0de2f7fd4ce

Publication date
2025

Document Version
Final published version

Citation (APA)
Yurrita Semperena, M. (2025). *Understanding Decision Subjects' Needs and Perceptions Towards Contestable AI Systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5792abf8-c8ff-4301-ba06-e0de2f7fd4ce

Understanding Decision Subjects' Needs and Perceptions Towards Contestable AI Systems

Mireia Yurrita Semperena

Understanding Decision Subjects' Needs and Perceptions Towards Contestable AI Systems

Understanding Decision Subjects' Needs and Perceptions Towards Contestable AI Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board of Doctorates,
to be defended publicly on May 12th 2025 at 15:00 o'clock.

by

Mireia YURRITA SEMPERENA

Master of Science in Industrial Engineering, Universidad de Navarra, Spain,
born in Beasain, Spain.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. ir. A. Bozzon,	Delft University of Technology, *promotor*
Prof. dr. S. C. Pont,	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. ir. G. J. P. M. Houben,	Delft University of Technology, Netherlands
Prof. dr. E. Giaccardi,	Politecnico di Milano, Italy
Prof. dr. ing. A. J. Klievink,	Leiden University, Netherlands
Prof. dr. ir. J. F. M. Masthoff,	Utrecht University, Netherlands
Prof. dr. P. A. Lloyd	Delft University of Technology, *reserve member*

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

Keywords: decision subjects, contestability, contestable AI, human-centered AI, trustworthy AI, needs, fairness perceptions, mixed-methods, human-computer interaction

Front & Back: Cover designed by Dr. Agathe Balayn and Dr. David Maxwell at a2d.studio. Cover typeset in Inter by Rasmus Andersson, under the terms of the SIL Open Font License.

ISBN 978-94-6518-043-4

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To those who felt powerless when contesting an unjust situation.
To those who made a conscious decision not to speak up.

Contents

Summary

Artificial Intelligence systems (AI) are increasingly being used for decision-making due to their capacity to process large amounts of data. While efficiency and lower decision-making costs make AI systems attractive tools for augmenting human decision-makers' capacities, many have highlighted that biases and inscrutability inherent to such systems could lead to harmful consequences for decision subjects. Contestability, i.e., a property that makes AI systems open to human intervention throughout their lifecycles, has been claimed to be essential for counteracting algorithmic harms. By enabling decision subjects to influence algorithmic outputs, contestable AI systems aim to safeguard decision subjects' rights to autonomy and dignity. Despite the interest and relevance of contestability in HCI, little is known about whether and how elements of contestable AI systems can empower decision subjects in algorithmic decision-making.

In this dissertation, we aim to generate empirical insights into decision subjects' needs for and fairness perceptions towards contestable AI systems in decision-making. By focusing on decision subjects, this dissertation leads to a set of recommendations for organizations setting up algorithmic decision-making processes. These recommendations encourage organizations to account for the interests of those impacted by algorithmic decisions from the early stages of the design process. To this end, we rely on a combination of qualitative and quantitative studies. We ground these studies in two high-risk decision-making contexts, one in the public sector –an illegal holiday rental identification scenario– and the other one in the private sector –a loan approval scenario. We conduct this dissertation in the context of the DCODE Marie Skłodowska-Curie Innovative Training Network (ITN).

In **chapter 2**, we review principles for *Trustworthy AI* and we locate contestability within the Trustworthy AI discourse. We synthesize the most prominent Trustworthy AI principles in a framework and operationalize such principles by breaking them down into criteria and manifestations. We advocate for opening up spaces for multi-stakeholder deliberation in AI design and assessment processes. The insights we generate as part of chapter 2 motivate the need to further look into contestability as a relevant trustworthy AI principle.

In **chapter 3**, we focus on contestability and look into decision subjects' information and procedural needs for meaningful contestability. We do not limit contestability to one-to-one appeal or recourse processes. This chapter presents an interview-based qualitative study. The findings in this chapter highlight the cooperative work behind contestability. Instead of conceiving the right to autonomy as individual self-determination, decision subjects' capacity for contestation is shaped by their interactions with actors (e.g., street-level buraucrats, third parties) involved in decision-making.

In **chapter 4**, we narrow contestability down into the *right to contest* automated decisions as interpreted from Article 22(3) of the European Union's General Data Protection Regulation. The elements that contribute to the right to contest automated decisions include explanations, human intervention and appeal mechanisms. Through a crowd-sourced quantitative study, we examine the interplay between explanations, human intervention (operationalized as human oversight) and appeal mechanisms, and evaluate their effect on decision subjects' fairness perceptions. The findings in this chapter show the positive effect of explanations and appeal mechanisms on informational and procedural fairness perceptions. The findings in this chapter also point towards the need to rethink traditional appeal mechanisms to capture the unique nature of AI systems in decision-making. The lack of empirical evidence of the effect of human intervention on decision subjects' fairness perceptions motivates the need to further investigate this element.

In **chapter 5**, we further look into the effect of human intervention on decision subjects' fairness perceptions. This chapter is motivated by the findings in chapter 4 and the relevance of human intervention within Article 22(3) of the GDPR. Through a mixed-methods study, we identify decision-makers' profile, model type and data provenance as three key properties that affect decision subjects' fairness perceptions towards human intervention. We also identify the Ability, Benevolence, and Integrity model as a useful tool for capturing perceptions towards decision-maker configurations that include varying levels of human intervention. The findings in this chapter confirm the positive effect of human intervention on decision subjects' fairness perceptions, but encourage an interpretation of human intervention that is not limited to human controllers making the final decision.

In **chapter 6**, we provide an overview of the research conducted as part of this dissertation. We discuss the implications of the research, provide a set of recommendations for organizations developing and deploying AI systems, for policy makers and for HCI researchers. We also reflect on our approach and highlight the limitations of this research.

Samenvatting

Kunstmatige intelligentiesystemen (AI) worden steeds vaker gebruikt voor besluitvorming vanwege hun capaciteit om grote hoeveelheden data te verwerken. Hoewel efficiëntie en lagere kosten AI-systemen aantrekkelijke hulpmiddelen maken om de capaciteiten van menselijke besluitvormers te vergroten, hebben velen benadrukt dat de vooroordelen en ondoorzichtigheid die inherent zijn aan dergelijke systemen schadelijke gevolgen kunnen hebben voor beslissingssubjecten. Betwistbaarheid, dat wil zeggen het openstellen van AI-systemen voor menselijke tussenkomst gedurende hun gehele levenscyclus, is mogelijk één van de sleutels om algoritmische schade tegen te gaan. Door beslissingssubjecten in staat te stellen algoritmische uitkomsten te beïnvloeden, streeft betwistbaarheid ernaar de rechten van beslissingssubjecten op autonomie en waardigheid te beschermen. Ondanks de interesse en relevantie van betwistbaarheid binnen HCI, is er weinig bekend over of en hoe betwistbare AI-systemen de invloed van beslissingssubjecten in algoritmische besluitvorming kunnen versterken.

In dit proefschrift streven we ernaar empirische inzichten te genereren over de behoeften van beslissingssubjecten aan betwistbare AI-systemen in besluitvorming, en hun eerlijkheidspercepties van diezelfde systemen. Door ons te richten op beslissingssubjecten, leidt dit proefschrift tot een reeks aanbevelingen voor organisaties die algoritmische besluitvormingsprocessen opzetten. Deze aanbevelingen moedigen organisaties aan om rekening te houden met de belangen van degenen die worden geraakt door algoritmische beslissingen vanaf de vroege stadia van het ontwerpproces. Hiervoor maken we gebruik van een combinatie van kwalitatieve en kwantitatieve studies. We baseren deze studies op twee contexten van besluitvorming met een hoog risico, één in de publieke sector - een scenario betreffende identificatie van illegale vakantieverhuur - en de andere in de private sector - een scenario betreffende beoordeling van leningen. We voeren dit proefschrift uit in de context van het DCODE Marie Skłodowska-Curie Innovative Training Network (ITN).

In **hoofdstuk 2** bespreken we principes voor betrouwbare AI en plaatsen we betwistbaarheid binnen het discours van betrouwbare AI. We maken een kader met een synthese van de meest prominente principes van betrouwbare AI en operationaliseren principes door ze op te splitsen in criteria en manifestaties. We pleiten voor het creëren van ruimte voor overleg tussen meerdere belanghebbenden in AI-ontwerp- en evaluatieprocessen. De inzichten die we genereren als onderdeel van hoofdstuk 2 motiveren de noodzaak om betwistbaarheid verder te onderzoeken als een relevant principe voor betrouwbare AI.

In **hoofdstuk 3** richten we ons op betwistbaarheid en kijken we naar de informatie- en procesbehoeften van beslissingssubjecten voor betekenisvolle betwistbaarheid. We beperken betwistbaarheid niet tot een-op-een beroeps- of bezwaarprocessen. Dit hoofd-

stuk presenteert op interviews gebaseerd kwalitatief onderzoek. De bevindingen in dit hoofdstuk benadrukken het coöperatieve karakter van betwistbaarheid. In plaats van het recht op autonomie op te vatten als individuele zelfbeschikking, wordt het vermogen van beslissingssubjecten om te betwisten gevormd door hun interacties met andere actoren (bijvoorbeeld bureaucraten op straatniveau, of derden) die betrokken zijn bij besluitvorming.

In **hoofdstuk 4** beperken we betwistbaarheid tot het recht om geautomatiseerde beslissingen te betwisten, zoals geïnterpreteerd vanuit artikel 22(3) van de Algemene Verordening Gegevensbescherming van de Europese Unie. De elementen die bijdragen aan het recht om geautomatiseerde beslissingen te betwisten, omvatten uitleg, menselijke tussenkomst en beroepsmechanismen. Door middel van een gecrowdsourcet kwantitatieve studie onderzoeken we de wisselwerking tussen uitleg, menselijke tussenkomst (geoperationaliseerd als menselijk toezicht) en beroepsmechanismen, en evalueren we hun effect op de percepties van eerlijkheid van beslissingssubjecten. De bevindingen in dit hoofdstuk tonen het positieve effect van uitleg en beroepsmechanismen op percepties van informationele en procedurele eerlijkheid. De bevindingen in dit hoofdstuk wijzen ook op de noodzaak om traditionele beroepsmechanismen te heroverwegen om de unieke aard van AI-systemen in besluitvorming vast te leggen. Het gebrek aan empirisch bewijs voor het effect van menselijke tussenkomst op de percepties van eerlijkheid van beslissingssubjecten motiveert de noodzaak om dit element verder te onderzoeken.

In **hoofdstuk 5** kijken we verder naar het effect van menselijke tussenkomst op de percepties van eerlijkheid van beslissingssubjecten. Dit hoofdstuk wordt gemotiveerd door de bevindingen in hoofdstuk 4 en de relevantie van menselijke tussenkomst binnen artikel 22(3) van de AVG. Door middel van een mixed-methods studie identificeren we het profiel van besluitvormers, het modeltype en de herkomst van gegevens als drie belangrijke eigenschappen die de percepties van eerlijkheid van beslissingssubjecten ten opzichte van menselijke tussenkomst beïnvloeden. We identificeren ook het Ability, Benevolence, and Integrity-model als een nuttig hulpmiddel voor het vastleggen van percepties ten opzichte van configuraties van besluitvormers die verschillende niveaus van menselijke tussenkomst omvatten. De bevindingen in dit hoofdstuk bevestigen het positieve effect van menselijke interventie op de perceptie van eerlijkheid van beslissingssubjecten, maar moedigen een interpretatie van menselijke interventie aan die niet beperkt is tot menselijke uitvoerders die de uiteindelijke beslissing nemen.

In **hoofdstuk 6** geven we een overzicht van het onderzoek dat is uitgevoerd als onderdeel van dit proefschrift. We bespreken de implicaties van het onderzoek, bieden een reeks aanbevelingen voor organisaties die betwistbare AI-systemen ontwikkelen en implementeren, voor beleidsmakers en voor HCI-onderzoekers. We reflecteren ook op onze aanpak en benadrukken de beperkingen van dit onderzoek.

Acknowledgements

This thesis would not have been possible without the direct or indirect contributions of many people who have accompanied me in my doctoral path. A big *thanks* to all of you.

To **Prof. Alessandro Bozzon** and **Prof. Sylvia C. Pont**, for giving me the freedom to explore my own research interests while offering critical feedback. To **Dr. Dave Murray-Rust**, for your dedication in the early stages of my doctoral studies.

To **Prof. Elisa Giaccardi**, **Prof. Judith Masthoff**, **Prof. Bram Klievink**, **Prof. Geert-Jan Houben** and **Prof. Peter Lloyd**, for accepting to be part of the committee and reviewing this thesis.

To my paranymphs, for bringing a little piece of the Basque Country to the Netherlands. To **Garoa Gomez Beldarrain**, for making me feel heard. Thanks for supporting me, even when you would have done things differently. To **Amaia Garmendia Urdalleta**, for countless evenings chatting in the kitchen. Our PhD topics are very different, yet we found similar difficulties in our paths. To **Garazi Muguruza Lasa**, for your clarity of thought. You are able to articulate how a PhD journey feels like few people can.

To the stellar senior PhD candidates (Doctors by now) that I have worked with, for your generosity. To **Agathe Balayn**, for being the co-author that I can consistently rely on. For helping me shape my thinking. And for always suggesting one more paper to look at. To **Tim Draws**, for your careful guidance. For highlighting the value of rigor, conciseness and attention to detail. Most importantly, thanks for reminding me there is plenty of life beyond academia. To **Kars Alfrink**, for always finding time for a chat. For bringing the philosophical touch to our conversations about contestability. And for insisting that the discussion section does not need to follow the same structure as the results section.

To my colleagues in the Knowledge and Intelligence Design team. To **Achilleas Psyllidis**, **Alejandra Gomez Ortega**, **Alice Vitali**, **Ariane Lucchini**, **Carlo van der Valk**, **Catalina Lagos**, **Céline Offerman**, **Di Yan**, **Evangelos Niforatos**, **Francesca Mauri**, **Gerd Kortuem**, **Hosana Morales**, **Jacky Bourgeois**, **James Broadhead**, **Jeff Love**, **Katherine Song**, **Roos Teeuwen**, **Samuel Freire**, **Sara Colombo**, **Shatha Degachi**, **Tianhao He**, **Tilman Dingler**, **Uğur Genç**, **Vasilis Milias**, **Wilfred van der Vegte** and **Wo Meijer**, for the support structures that formally and informally we created to help each other. For our very active Mattermost channel. Special thanks to **Himanshu Verma**, for transmitting your passion for imaginaries, shared cognition and statistics. Because there is nothing more exciting than finding mediation effects in our data. Even if I then visualize them using Times New Roman font.

To my colleagues in Studio Lab (and beyond). To **Willem van der Maden**, for being the coolest DJ in the office. Thanks for opening the doors of your home in Copenhagen. To **Wasabii Ng**, for stressing that healing takes time. To **Paula Hueso**, for making all nostalgic Spaniards in the Netherlands dance to the rhythm of Paquito El Chocolatero. To **Caiseal Beardow**, **Sofie-Amalie Dideriksen**, and **Marie-Therese Sekwenz**, for those drinks that we so much struggle to organize.

To my colleagues in the DCODE network. To **Sonja Rattay**, for your sweet soul. Thanks for regularly checking on me. To **Mugdha Patil**, for engaging in the demanding task of finding the best pizza place in Torino. To **Irina Shklovski**, for challenging my assumptions and for welcoming me in the Human-Centered Computing team at KU. To **Shalini Kurapati** and **Luca Gilli**, for your help and kindness during my secondment at Clearbox AI.

To my fellow PhDs in the Web Information Systems group. To **Garrett Allen** and **David Maxwell**, for the ciders that we enjoyed in bars all over Delft. For our shared frustrations. For bearing (and trading) with me during Catan games.

To **Gerard Ribera** and **Ángel de Castro**, for welcoming me in Choorstraat. Thanks for helping me settle in the Netherlands and for making the beginning of my new life so smooth and easy. To the gang of Spaniard, Catalan, Belgian, French, Argentinian and Mexicans in Delft. To **Andrea Palet**, **Antonin Bontempelli**, **Jorge Baeza**, **Judith Cueto**, **Julia Juan**, **Manuela García**, **Marcela Aragón**, **Mario Aragonés**, **Martina Benito**, **Miquel Piris**, **Miriam Cañones**, **Ricard Mallafre**, **Sara Vienne**, **Victor Arribas**, **Victor Sánchez**. For the international weddings, summer barbecues, birthday parties, and random weekend drinks. Thanks for providing a space where I could unplug from my PhD.

To **Andoni Pagola** and **Oihane Mujika**, for saying *yes* to every crazy idea I suggest. Our trips across Turkey, Jordan, Peru, Colombia and Japan have given me the motivation to push for deadlines, to accept that good enough is actually enough, and to click the *submit* button.

To my parents, sister, brother-in-law and nephew. **Ama, aita**, eskerrik asko zuen alabei etorkizun oparo bat ematearren egindako sakrifizio guztiengatik. Betidanik, nire kuriositatea asetzera bultzatu nauzue. Eskatzen dena baino pixka bat haratago heltzera. Asmo handiko helburuak ezartzera. Jakinmin, esfortzu eta anbizio horrek ahalbidetu ditu nire orain arteko lorpen asko. Azken urte honetan, ordea, noiz gelditu jakitearen garrantzia gogorarazi didazue. Mila esker lezio baliotsu honegatik ere. **Enara** eta **Eneko**ri, edozertarako telefonoaren beste aldean egoteagatik. Nahiz eta kilometro askotara egon, gertu sentiarazteagatik. Bueltatzen naizen bakoitzean nire presentzia ospatzeagatik. Gure **Iakes** txikiari. Etxeko bizipoza zaren horri. Mundua deskubritzen nekaezin zabiltzan honetan, jakin izeba beti izango duzula alboan.

To **Lorenzo** (a.k.a. **Lontxo** para la parte Euskaldun de la familia). Gracias por tu apoyo diario. Por aportar serenidad y templanza a mi impaciente y enérgica personalidad. Por saber que no hay mal día que un vinito blanco y una buena *amatriciana* no puedan solucionar. Gracias también por ser la persona que más veces me dice *no.* Que sepas que contribuyes a la mejora de mis dotes argumentativas y me ayudas a ser un poquito más *paraculo* cada día. Por muchos años más a tu lado. Preferiblemente con vistas a La Concha.

Terminology

Term	Definition
Algorithmic decision-making	Decision-making processes that are driven or augmented by algorithmic systems.
Algorithmic systems	Systems that are driven by a set of instructions (models) that are used to solve a problem. I adopt a sociotechnical perspective to this definition and, therefore, account for both human and machine actors involved in the functioning of algorithmic systems.
Artificial Intelligence systems	Computational systems that are designed for interpreting external data, processing that data, and using those results for performing specific tasks [212]. AI systems can be stochastic, deterministic, or a combination of both. I adopt a sociotechnical perspective to this definition and, therefore, account for both human and machine actors involved in the functioning of AI systems.
Appeal process	Act of opposing an algorithmic decision because it is considered to be faulty, incorrect or incomplete [412]. The scope is limited to the output of the system.
Contestability (by design)	Quality that makes an algorithmic system open and responsive to human intervention throughout its lifecycle [9]. The scope is not limited to the output of the system. It allows to contest design choices early in the development process [14].
Decision subject	Individual affected by an algorithmic decision [386].
Distributive justice	Dimension of justice related to the outcome resulting from a decision-making process [95].
Fairness perceptions	Measurements of appropriateness that evoke different dimensions of justice [94].
Human controller	Domain expert interacting with the algorithmic system so as to ensure the quality of the algorithmic decision (i.e., algorithmic decision overseen by a human [416]) or to make a decision themselves (i.e., human decision-making with algorithmic elements [416]) [9].
Human intervention	Mediation by a human being who has the competence and authority to possibly change the decision made by an AI system [131].
Human oversight	Feature of a hybrid decision-making configuration that indicates a level of intervention where a human identifies and corrects potential mistakes made by an algorithmic system [14].
Human reviewer	Actor responsible for evaluating the correctness of the algorithmic decision during the contestation process [271].
Hybrid decision-making	Decision-making processes led by algorithmic systems where a human intervenes to identify and correct potential mistakes made by the algorithmic system [14].
Informational justice	Dimension of justice related to the information provided to decision subjects about a decision-making process [95].

Interactional justice	Dimension of justice related to the treatment received by decision subjects as part of a decision-making process [95].
Justice	Multi-dimensional construct that studies criteria for determining what is appropriate or correct [94].
Machine Learning driven systems	Subdivision of Artificial Intelligence systems. Machine Learning driven systems are stochastic in nature.
Practitioner	Professionals involved in the development, deployment or governance of Artificial Intelligence systems
Procedural justice	Dimension of justice related to the nature of the process that leads to decisions [95].
Recourse	Act of modifying the input variables of an algorithm or Artificial Intelligence model to change the output [404].
Trustworthy Artificial Intelligence	Artificial Intelligence systems that are lawful, ethical, and robust [140].

Table 1: Summary of key terms and corresponding definitions.

1

Introduction

1.1. Context and Motivation

In recent years, the usage of algorithmic systems based on Artificial Intelligence (AI) in general, and more specifically, on Machine Learning (ML) have become ubiquitous in several different domains. In the public sector, for example, AI systems are pervasive in the detection of fraud [200, 445], in public employment and medical services [367, 385, 428], or in policing [69, 265, 285]. In Europe, some initial analyses indicate that the Netherlands represents the country with the highest number of public initiatives where AI technologies are introduced [293, 309]. Artificial Intelligence has even been designated the status of "key technology" [338] for its prospects to bring individual, social, and environmental benefits[1] [140]. The rapid adoption of this technology is justified under its potential to increase efficiency, lower decision-making costs [230], and provide individual citizens with better outcomes [339]. This is due to AI's capacity to consistently process big amounts of data [230].

While many have highlighted the potential benefits of incorporating artificial skills to transform existing workflows [39], the application areas in which these systems are introduced is an important factor to consider. Most of the aforementioned application areas directly impact citizens' safety and fundamental rights and are considered *high-risk* areas by recent regulatory efforts such as the European Union's Artificial Intelligence Act (EU AI Act) [133]. Concerns about vulnerability [199], bias [74], and inscrutability [25] of AI systems in such applications, hence, become especially problematic. The childcare benefits scandal in the Netherlands (colored box below) is an illustration of it.[2]

The childcare benefits scandal of 2018 in the Netherlands [19] is an illustration of the harmful consequences of adopting biased AI systems with a black-box configuration for high-risk public decision-making. A few years earlier, the Dutch government had introduced a self-learning risk classification model to create risk profiles of childcare benefits applicants for a rapid identification of individuals who were likely to commit fraud. Due to the opacity of the system, affected applicants were deprived of meaningful information to understand their individual situation. This led to thousands of benefits applicants being falsely accused of fraud, while making it impossible for such applicants to oppose, correct or remedy their situation.

In an effort to mitigate the potentially harmful consequences of adopting AI systems for high-risk decision-making, several actors, such as the European High-Level Expert Group, have emphasized the need to support the design, deployment, and governance

[1] While the negative environmental impact of AI models has been much debated [266], AI is also believed to be a powerful tool that can help reduce greenhouse gas emissions by e.g., optimizing central heating systems to minimize their environmental impacts [342]

[2] Information concerning the Dutch Childcare Benefits Scandal included in this dissertation is based on the 2021 report *Xenophobic Machines: Discrimination Through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal* by Amnesty International [19] https://www.amnesty.nl/content/uploads/2021/10/20211014_FINAL_Xenophobic-Machines.pdf?x25337. At the time of writing this dissertation (September 2024) the legal procedures dealing with the scandal are still ongoing. We encourage the interested reader to follow the latest events through reliable sources.

of AI systems through socio-political deliberation [140]. To ensure AI systems are trustworthy, several ethical principles have been defined (e.g., privacy, fairness, explainability, contestability) [138]. In this dissertation, we focus on the principle of *contestability*.

Contestability has a multifaceted nature and fits different definitions. In the field of AI ethics, contestability has been defined as a core principle for trustworthy AI, which contributes to addressing the politics and power imbalances behind the implementation of AI systems [269, 406]. In the legal realm, contestability has been claimed to be a fundamental democratic right [269]: a means for empowering decision subjects to influence decisions that can impact their lives. By enabling influence over algorithmic outputs, contestability safeguards decision subjects' rights to autonomy and dignity, and represents a way to counteract harmful algorithmic decisions [9, 14, 269]. Contestability, has been, therefore, claimed to foster procedural justice and to positively contribute to decision subjects' fairness perceptions in algorithmic decision-making [9]. Scholars in the field of human-computer interaction (HCI), instead, define contestability as a system property that makes AI systems open and responsive to human intervention throughout their lifecycles [9]. AI systems that embed contestability as a "deep system property" [405] (i.e., contestability by design [14]), are identified as *contestable AI systems* [9]. When *designing for* contestability, values that go into the design of AI systems (e.g., values that condition decisions about which data to include, which data to exclude, or what "good" data is) are surfaced [9, 405]. Surfacing those values, in turn, contributes to holding organizations developing, deploying, and adopting algorithmic decision-making processes accountable.

In the case of the Dutch childcare benefits scandal, making the AI system contestable, would have meant e.g., questioning the appropriateness of using a risk classification model at the early stages of its design, implementing mechanisms to explain the decisions to affected applicants, or providing information about the workings of the model to enable scrutiny by oversight bodies.

Even if the field of contestable AI has generated considerable interest in HCI in the last years, this research area is still nascent. Many of the guidelines on how to design for contestability are conceptual [9, 179, 269] and there is little empirical insight into how contestable AI can empower decision subjects in algorithmic decision-making. The relevance and scarcity of empirical evidence on how to design and deploy contestable AI systems demonstrates the need for more research in this area. In this dissertation, we address this need. The goal of this dissertation is *to generate recommendations for contestable AI design and deployment* so that organizations setting up algorithmic decision-making processes shape these systems in a way that effectively empowers decision subjects to influence algorithmic outputs. We do so, by *generating empirical insights* into decision subjects' needs for and fairness perceptions towards contestability in algorithmic decision-making configurations. This dissertation, therefore, answers to the following overarching research question: **how can decision subjects' needs and fairness perceptions be considered to inform the development and deployment of contestable Artificial Intelligence systems for algorithmic decision-making?**

Our focus on decision subjects' needs and fairness perceptions to inform contestable

AI design and deployment is motivated by the fact that algorithmic decision-making processes may have both economic and socio-emotional consequences for decision subjects [96]. By capturing the concerns and needs of individuals and communities that are (or will most likely be) impacted by the use of AI systems, we make sure that their interests are considered from the very early stages of the AI design process [251]. This is a first step towards avoiding many of the harmful effects of implementing AI systems in decision-making processes [377] and ensuring algorithmic decision-making is perceived as an acceptable and legitimate alternative to human-led decision-making [95, 162, 253, 259, 393, 400]. Our research questions and methodological choices are framed so as to capture the needs for and perceptions towards contestable AI of those subject to algorithmic decisions. We study these needs and perceptions in two different high-risk decision-making contexts; one in the private sector (i.e., a loan approval scenario) and one in the public sector (i.e., an illegal holiday rental identification scenario).

This dissertation has been developed as part of a broader European project, the DCODE Network [306], where we aim to reconcile human values and algorithmic logic. Within the objectives of the DCODE Network [306], this dissertation focuses on anticipating desired interactions between humans and AI systems for supporting sustainable digital futures. In line with the aims of the project, we adopted a human-centered approach for the principled design of contestable AI.

This dissertation is structured as follows. We first locate contestability within the current trustworthy AI discourse and motivate the need to further look into contestability as a prominent ethical principle. Through a combination of qualitative and quantitative approaches, we then explore decision subjects' broader needs for contestability. We next study decision subjects' fairness perceptions towards the *right to contest* automated decisions as interpreted from Article 22(3) of the European Union's General Data Protection Regulation (GDPR) [131]. We finally further dive into the role of human intervention in shaping decision subjects' fairness perceptions. Human intervention is defined by the GDPR [131] as a safeguard against automated decision-making that conditions the right to contest automated decisions [416]. We, therefore, contribute to the current literature on contestable AI with a series of empirical insights into decision subjects' needs for and fairness perceptions towards contestability in algorithmic decision-making. We inform the development and deployment of contestable AI by discussing the implications for practice and for research of the generated empirical insights and by providing stakeholder-specific recommendations to meaningfully empower decision subjects in algorithmic decision-making.

1.2. Research Landscape and Knowledge Gaps

In this section, we provide an overview of the research landscape where this dissertation is located. The research landscape includes topics on (1) trustworthy AI principles, (2) design of contestable AI systems, (3) the right to contest automated decisions, and (4) human intervention in algorithmic decision-making. In each section, we highlight the four knowledge gaps that this dissertation addresses within that research landscape (see Figure 1.1).

1.2.1. Principles for Trustworthy Artificial Intelligence

In view of the potentially harmful consequences that Artificial Intelligence systems could perpetuate if not adopted responsibly, there has been a proliferation of guidelines to ensure that AI systems are trustworthy. Since 2016, tens of organizations both in the public (e.g., European Commission, U.S. National Science and Technology Council, Standards Administration of China) and private (e.g., Microsoft, Google, IBM, Telefonica) sphere have defined high-level ethical principles that aim at guiding the design, deployment, and governance of AI systems [296, 298]. Public institutions like the European Commission, through their High-Level Expert Group (HLEG) [140], for example, defined Trustworthy Artificial Intelligence as an approach to developing AI that ensures that such systems are (1) lawful (i.e., they respect policies and regulations that are in place), (2) ethical (i.e., they are aligned with societal values and ethical principles), and (3) robust (i.e, the models that compose these systems are technically insensitive to misperformance or miscalculations [395]). The EU's HLEG additionally defined seven key requirements that trustworthy AI should respect: (1) human agency and oversight, (2) robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and (7) accountability [140]. Private companies like Microsoft, instead, defined six principles as part of their responsible AI standard: (1) accountability, (2) transparency, (3) fairness, (4) reliability and safety, (5) privacy and security, and (6) inclusiveness.

Even if standards suggested by public and private organizations slightly differ in prioritizing certain principles over others, approaches to trustworthy AI have seemingly converged into a set of principles [296]. Such convergence is visible in an analysis conducted by Fjeld et al. [138]. Fjeld et al. [138] studied thirty-six AI ethics documents collected through a purposive sampling method and identified eight common themes. These themes include (1) privacy, (2) accountability, (3) safety and security, (4) transparency and explainability, (5) fairness and non-discrimination, (6) human control of technology, (7) professional responsibility, and (8) promotion of human values. According to the authors, the most recent documents cover all main themes identified through their analysis. The convergence of AI ethical documents into these eight themes opens up the possibility to the emergence of a "normative core" that could guide a principle-based approach to the design, deployment, and governance of Artificial Intelligence systems [138].

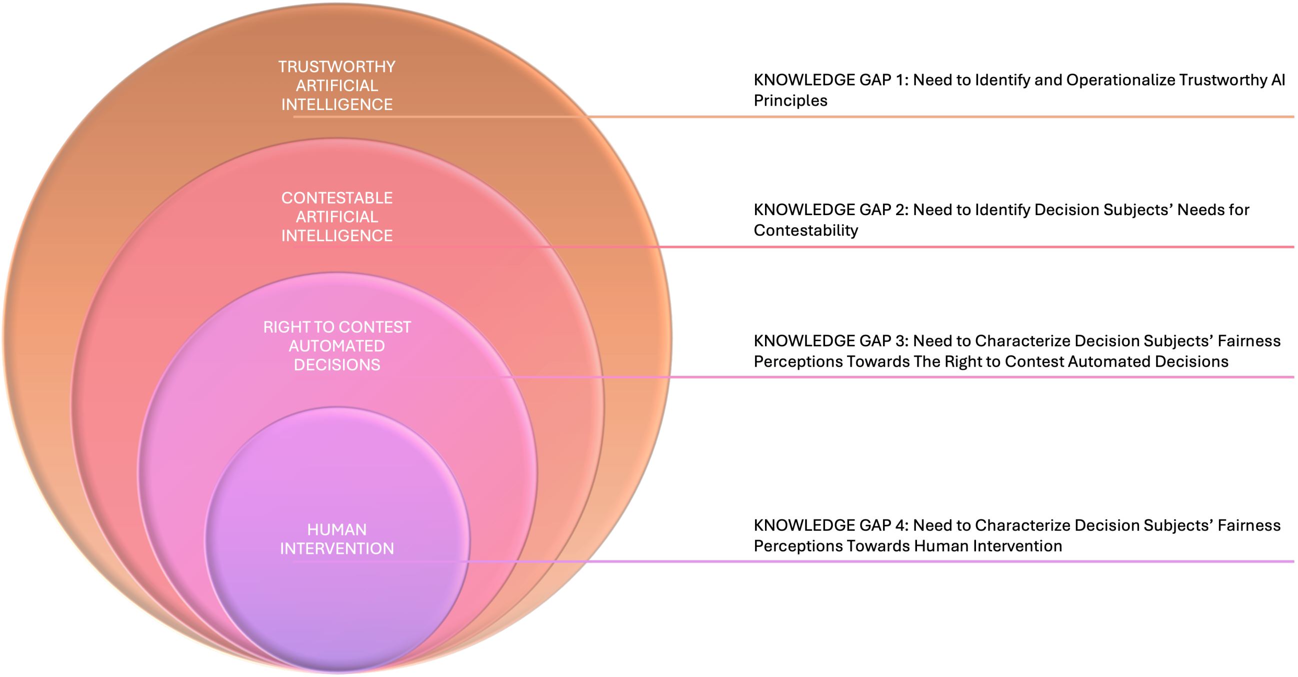

TRUSTWORTHY ARTIFICIAL INTELLIGENCE
CONTESTABLE ARTIFICIAL INTELLIGENCE
RIGHT TO CONTEST AUTOMATED DECISIONS
HUMAN INTERVENTION
KNOWLEDGE GAP 1: Need to Identify and Operationalize Trustworthy AI Principles
KNOWLEDGE GAP 2: Need to Identify Decision Subjects’ Needs for Contestability
KNOWLEDGE GAP 3: Need to Characterize Decision Subjects’ Fairness Perceptions Towards The Right to Contest Automated Decisions
KNOWLEDGE GAP 4: Need to Characterize Decision Subjects’ Fairness Perceptions Towards Human Intervention

Figure 1.1: Knowledge gaps addressed in this dissertation.

Knowledge Gap 1: Need to Identify and Operationalize Trustworthy AI Principles

To address the overarching research question of the present dissertation, we start by looking into trustworthy AI principles and their operationalization. The ultimate goal of this first step is to locate contestability within the current discourse of trustworthy AI and identify guidelines for operationalizing it.

Initiatives to define trustworthy AI are seemingly converging into a set of common principles [296]. However, these principles have been criticized for being too high-level, and limited to vague value statements [296]. Even if these trustworthy AI initiatives claim to be action-guiding, they provide few practical and actionable recommendations that those desigining, deploying, and governing AI systems can put into practice [169]. For example, Fjeld et al. [138] found that the principle of human control was present in 69% of the reviewed standards. However, the way in which meaningful human control can be operationalized to address responsibility gaps when interacting with AI systems [81] is not straightforward. Documents dealing with principles for trustworthy AI seldom unpack and operationalize those principles.

Furthermore, fundamental tensions inherent in principles such as *fairness* are not discussed in depth in trustworthy AI standards [169]. If we define fairness as “equalized odds”, for a classification task driven by an AI system, a *fair* model would mean that both the true and false positive rates should be equal for the protected and unprotected groups [284]. Instead, if we define fairness as “equal opportunity”, only the true positive rates should be equal for the protected and unprotected groups [284]. Work in fairness in Machine Learning have characterized *fairness* through up to ten different definitions, some of which are mutually exclusive. Documents dealing with principles for trustworthy AI, nonetheless, do normally not dive into such tensions. These documents provide little insight into how discussions around the tensions arising from the design of socio-technical systems can take place in practice.

There is, therefore, a need to operationalize high-level ethical principles for trustworthy Artificial Intelligence, while opening up spaces for multi-stakeholder deliberation in the design and assessment of socio-technical systems.

1.2.2. Contestable Artificial Intelligence

Contestability has been described as a prominent ethical principle by many trustworthy AI standards (e.g., [10, 31, 132, 138, 209, 248, 269, 391]), key to ensuring trustworthy AI. In this dissertation, we use the term *contestable AI* [9] to refer to AI systems that uphold the principle of contestability by design. We use the term *contestable algorithmic decision-making* to refer to algorithmic decision-making processes that rely on contestable AI systems. Contestability has been defined as the ability to oppose an action, either because the action is perceived as mistaken or simply wrong [10]. Contestability aims to safeguard decision subjects’ rights to dignity and autonomy [9]. Contestability has been conceptualized in various ways [269]. *Recourse* refers to the ability of a person to “change the decision of the model through actionable input variables” [404]. In order for a decision subject to be able to exercise recourse, they need an actionable set of factors or counterfactual scenarios to obtain a desired outcome [206, 410]. When contestability is characterized as the act of incorporating *appeal mechanisms* in decision-making, there is no assumption that the original decision was correct in the very first place [412].

Appeal mechanisms give decision subjects the ability to correct faulty decisions [412]. Some authors consider that contestability is not limited to recourse and appeal mechanisms [269] and define the concept of *contestability by design* [9, 14, 355]. Contestability by design complements the ability to contest an outcome with the ability to iterate on the decision-making process [407]. Contestability by design has also been defined as a mechanism that enables AI systems to be "open and responsive to human intervention throughout the system lifecycle" [9].

The nature and purpose of contestability are open to dispute [269]. On the one hand, some claim that contestability is an end in itself; a fundamental principle to democracy. On the other hand, contestability has also been defined as a means to supporting principles such as fairness or accountability. Legal psychology scholars [253, 393] have traditionally framed contestability as a form of procedural justice; an instrument to ensure that decision subjects are empowered to express their views during a decision-making process. Contestability is claimed to increase decision subjects' fairness perceptions towards decision-making processes [9]. Vaccaro et al. [405], instead, described *designing for* contestability as a mechanism to surface values embedded in the design of AI systems; a way to ensure AI systems are fair, accountable and trustworthy.

A Multi-Stakeholder Conceptualization of Contestability. The field of contestable AI is nascent but is rapidly growing in the last years. Prior research (e.g., [9, 14, 269, 355]) has set the theoretical foundations for conceptualizing contestability. In a prominent recent work, Alfrink et al. [9] suggested a framework to characterize contestable AI. This framework consists of five socio-technical features and six practices that contribute to contestable AI. Features towards contestable AI include built-in safeguards against harmful behavior (e.g., adversarial secondary systems for decision-making), interactive controls to influence automated decisions, explanations about algorithmic behavior, human review and intervention requests, and tools for third parties and decision subjects to scrutinize AI systems. Practices contributing to contestable AI include ex-ante safeguards (e.g., early-stage assessments to prevent harms), agonistic approaches to AI development, quality assurance during development and after deployment of AI systems, risk mitigation strategies, and third party oversight.

Alfrink et al.'s [9] conceptualization of contestability is multi-stakeholder in nature, i.e., features and practices contributing to contestability involve a number of different stakeholders (see Figure 1.2). For example, interactive controls over automated decisions are mainly directed at domain experts interacting with AI systems to make decisions (i.e., human controllers). Instead, explanations of how the system behaves can be directed at both human controllers —to identify potentially erroneous outputs from the system— or at decision subjects —to exercise their right to contest automated decisions. Practices that contribute to contestability, such as quality assurance during AI system development, fall under the responsibility of organizations developing or deploying AI systems.

Knowledge Gap 2: Need to Identify Decision Subjects' Needs for Contestability

Despite the importance of contestability for trustworthy AI, most work in contestable AI is conceptual in nature (e.g., [9, 269]) and its operationalization has received comparatively little attention. For ensuring AI systems are trustworthy in practice, recent work

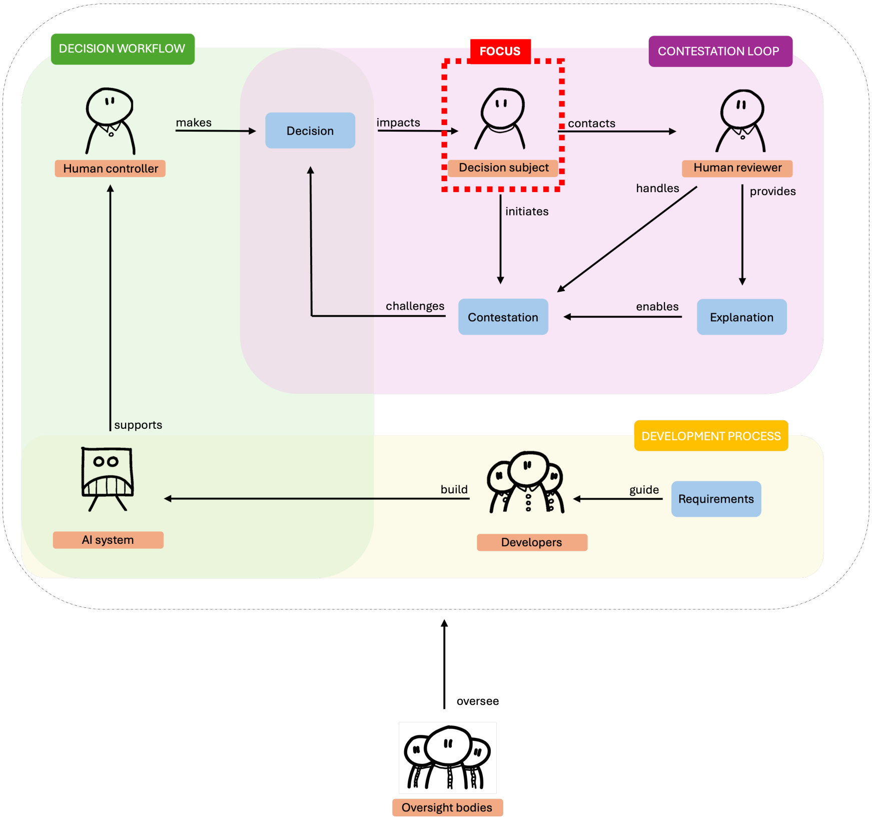


Figure 1.2: Overview of the actors and processes involved in contestable AI (adapted from [7, 9]). Contestable AI systems are open and responsive to human intervention throughout their lifecycles, i.e., throughout the development process, decision-making workflow and contestation loop. The focus of this dissertation is on characterizing *decision subjects'* needs for and perceptions towards contestability.

in the field of human-computer interaction (e.g., [12, 70]) has advocated for adopting a human-centered approach; by capturing the concerns and needs of individuals and communities that are most likely to be impacted by algorithmic decisions. The objective of such an approach is to generate AI design and deployment guidelines that ensure organizations responsible for algorithmic decision-making shape these processes in a way that effectively empowers impacted decision subjects and avoids unwanted effects for them. This is necessary for aligning algorithmic decision-making processes with decision subjects' perspectives and standards of justice. In the field of contestable AI, few studies have focused on operationalizing contestability as exercised by decision subjects.

One of the few studies focusing on decision subjects' needs for contestability was conducted by Vaccaro et al. [407]. Vaccaro et al. [407] conducted several participatory workshops with communities impacted by content moderation mechanisms and identified their procedural needs for appeal mechanisms. Through these workshops Vaccaro et al. [407] identified three major needs: (1) the need for direct or indirect forms of representation, (2) the need for support mechanisms for users to communicate with the platforms, and (3) the need for treating decision subjects in a compassionate way. Whether these procedural needs are applicable to contestability (beyond one-to-one appeal mechanisms) in high-risk decision-making contexts (beyond content moderation) is still unclear.

Additional to procedural needs for contestability, it should be noted that information about the decision-making is an essential step for decision subjects to be able to contest algorithmic decisions [269]. Hénin and Le Metayer [179] argue that this information should be given in the form of *justifications*, i.e., information that demonstrates the appropriateness of the algorithmic decisions. To date, most of the work dealing with decision subjects' information needs for contestability has been limited to recourse. Information for recourse normally concerns counterfactual explanations that decision subjects can act upon to change the output of the AI system (e.g., [206, 410]). Decision subjects' information needs for meaningful contestability (beyond recourse) are still to be explored.

In view of the above, guidelines on how to develop and deploy contestable AI systems, might not be aligned with decision subjects' needs for meaningful contestability. This might lead to contestability not effectively empowering decision subjects who are faced with decisions that have both economic and socio-emotional consequences for them [95]. There is, therefore, a need to identify decision subjects' procedural and informational needs for meaningful contestability beyond recourse or appeal.

1.2.3. The *Right to Contest* Automated Decisions

Beyond the fields of (1) fairness, accountability, and transparency in AI and (2) human-computer interaction, recent policy efforts, such as the European Union's General Data Protection Regulation (GDPR) [131], have also acknowledged the need for contestability in algorithmic decision-making. Article 22 of the GDPR deals with algorithmic decisions that are based *exclusively* on automated decision-making, i.e., without any *human intervention* with competence and authority to change the outcome of the system, if needed. Article 22(1) acknowledges the right of the data subject [3] not to be subject to fully automated decision-making when it produces legal effects or similarly significant effects on him or her. Article 22(2), however, enumerates three exceptions to the rule. Decisions based exclusively on automated decision-making are allowed if (a) it is necessary for entering into a contract between the data subject and the data controller, (b) it is authorized by the laws of the Member State that the controller is subject to, or (c) the data subject gives explicit consent to it. For cases where automated decision-making is permitted, Article 22(3) states that:

> *"In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the deci-*

[3] When referring to the GDPR [131], we will mimic their terminology and use *data subject* and *data controller* to refer to *decision subject* and *human controller*, respectively.

sion.”

There is an open debate among legal scholars around the interpretation of Article 22[4] (e.g., [287, 355]). Some of the points raised in such debates are of interest for this dissertation. As summarized by Sarra [355], there are four points concerning the nature of a contestation and the existence of the right to contest that are implicit in Article 22(3) and that are worth mentioning.

First, Article 22(3) shall not be interpreted as the right of the data subject to an effective remedy before a tribunal. This is already acknowledged as a fundamental right in the European Union, which would make the provision in Article 22(3) redundant. Instead, Article 22(3) shall be interpreted as the requirement for controllers and their organizations to create specific contestation structures *within* the organization. This shall allow data subjects to engage in a dialectical exchange within the framework of that organization [354].

Second, a contestation is not the mere expression of a personal opinion. A contestation is a *defensive* act that requires an argumentative effort from the data subject in order to challenge the automated decision [355]. For engaging in such an argumentative effort, data subjects need arguments. To build arguments, functional and meaningful explanations [368] are required. This interpretation, therefore, frames the much-debated *right to explanation* as a pre-condition for the right to contest an automated decision. The level of detail of explanations should be such that enables data subjects to engage in a meaningful contestation [355].

Third, the safeguard of *human intervention* conditions the existence of decision-subjects’ *right to contest*. Article 22(3) is only applicable in cases where the decision-making is fully automated. The act of asking for human intervention, therefore, exhausts the right to contest such decision (i.e., human intervention makes the decision-making not fully automated) [355]. In other words, human intervention shapes the very existence of the right to contest automated decisions.

Fourth, human intervention might be interpreted as the only safeguard that could prevent a chain of automatisms from being triggered when data subjects exercise their right to contest automated decisions. Article 22(3) does not limit the mechanisms that data controllers could use to deal with a contestation. Data controllers (or their organizations) could, therefore, set up another automated procedure for dealing with disputes [354]. In such a case, Article 22(3) would still apply for the contestation process. Requesting human intervention would, therefore, be the only way of interrupting the chain of automatisms. This would, in turn, exhaust decision subjects’ right to contest such automated decision.

[4] Discussing and comparing different legal interpretations of Article 22 of the GDPR is out of the scope of this dissertation. We briefly point to some of the aspects discussed in such legal interpretations as a way to justify the theoretical and design choices made in the remaining of this dissertation. Even if inspired by legal interpretations, this dissertation mostly contributes to the field of human-computer interaction.

Knowledge Gap 3: Need to Characterize Decision Subjects' Fairness Perceptions Towards the *Right to Contest*

Despite claims about the potential of contestability to safeguard decision subjects' legitimate interests [131] and improve their fairness perceptions in algorithmic decision-making [9], to the best of our knowledge, there is no empirical evidence that supports such claim. To characterize the effect of contestability on decision subjects' fairness perceptions, some studies have systematically evaluated different (existing) approaches to contestability. Vaccaro et al. [406], for instance, looked into the effect of different appeal processes on decision subjects' fairness perceptions in a content moderation scenario. Participants were assigned one of the following types of appeal: (1) no option to appeal the decision, (2) option to appeal the decision in written format to a human, (3) option to appeal the decision in a written format to an algorithm, or (4) option to appeal the decision to an algorithm through behavioral change. Vaccaro et al. [406] found that none of the appeal designs presented to participants improved their perceptions of fairness, accountability and trustworthiness.

From the design of their study, it can be seen that Vaccaro et al. [406] studied the effect of appeal mechanisms on decision subjects' fairness perceptions in a vacuum. However, to evaluate and validate approaches to contestability as suggested by policy efforts like the GDPR [131], it is necessary to account for the intricate entanglements among the elements that constitute contestation processes. As mentioned in section 1.2.3, when interpreting the right to contest automated decisions from Article 22(3) of the GDPR [131], it becomes evident that appeal mechanisms are shaped by the existence of explanations and human intervention [355]. Explanations, human intervention, and appeal mechanisms, therefore, co-shape decision subjects' procedural rights in the decision-making process [355]. Explanations, human intervention, and appeal mechanisms might, similarly, also co-mediate decision subjects' fairness perceptions. When evaluating approaches to contestability, like the *right to contest* automated decisions as interpreted from the GDPR [131], it is, therefore, necessary to capture the individual *and* combined effects of those three elements (i.e., explanations, human intervention, appeal mechanisms) on fairness perceptions.

1.2.4. Human Intervention in Algorithmic Decision-Making

The European Union's General Data Protection Regulation [131] refers to *human intervention* as the *intervention by a human being who has the competence and authority to possibly change the decision made by an AI system in a decision-making process* [354]. Based on the previous definition, it could be argued that only the intervention of the human controller involved at the end of the decision-making qualifies as "human intervention" according to the GDPR [131]. Human intervention is one of the three safeguards defined by Article 22(3) of the GDPR [131] to ensure that the "data subject's rights and freedoms and legitimate interests" [131] are preserved when decisions are made by exclusively relying on automated decision-making. There are, however, some limitations that might hinder human controllers' capacity to consider decision subjects' rights, freedoms and interests.

First, in an era where several of the AI systems used for decision-making are opaque and non-interpretable [346], the capacity of the human controller to effectively identify the cases in which the output of the system needs to be changed is highly questionable. In the field of human-computer interaction, it is well known that human controllers often suffer from cognitive biases and might end up overrelying on AI systems [72, 174, 198]. Such phenomenon might lead to human intervention boiling down to a simple confirmation mechanism of automated decisions [354]. This would directly affect the capacity of human intervention to provide decision subjects with any meaningful protection.

Second, even if the human controller were in a position to intervene and change the output of the system, this does not necessarily mean that they would do it to attend the needs of the decision subject that requested such intervention. The human controller might change the output of the system, e.g., following the interests of the company where they are embedded. No restrictions in Article 22(3) would prevent this from happening [354]. Since Article 22(3) is applicable only in the cases where the decision-making is fully automated, calling for a human intervention —without necessarily attending to the decision subject's needs— would, in turn, condition the very existence of decision subjects' right to contest an automated decision [355] (see section 1.2.3).

From the above reasoning, it can be concluded that (1) human intervention might exhaust the right to contest automated decisions and, at the same time, (2) is no guarantee that decision subjects' needs will be considered [354].

Knowledge Gap 4: Need to Characterize Decision Subjects' Fairness Perceptions Towards Human Intervention

From the interpretation of Article 22(3) of the GDPR [131], it can be inferred that human intervention conditions decision subjects'. It is, therefore, important to understand how human intervention affects decision subjects' fairness perceptions. However, studies looking into the effect of human intervention on decision subjects' fairness perceptions have led to ambiguous and oftentimes contradictory results [382]. The effect of human intervention has been mostly studied by comparing decision subjects' fairness perceptions towards human-led decision-making *vs.* algorithmic decision-making. In most cases, humans are perceived to be fairer than algorithms [80, 247, 264]. However, the lack of consistency across empirical studies suggests that fairness perceptions towards different decision-making configurations are highly context-dependent [382]. Lee et al. [247], for instance, compared perceptions of fairness towards human-led *vs.* algorithmic decision-makers in tasks requiring human and mechanical skills. The authors found that for tasks requiring human skills, algorithmic decision-making processes were perceived as less fair than human-led processes. In tasks requiring mechanical skills, instead, both algorithmic and human-led decision-making processes were perceived as equally fair. For healthcare, Longoni et al. [264] found resistance towards automated healthcare providers. The reason for this was mainly the incapacity of Artificial Intelligence systems to account for the uniqueness of each consumer. Wang [417] conducted a study to capture fairness perceptions in bail hearings and found that participants were more willing to accept discriminatory decisions when the decision-making was led by AI systems rather than humans.

While most literature in the research area of human-computer interaction has focused on comparing fully human-led to fully algorithmic decision-making processes —

in binary terms—, a comparatively smaller research effort has been devoted to additionally studying hybrid decision-making configurations (i.e., algorithmic decision-making where there is some level of human intervention) [382]. Even if combining human and artificial skills has been claimed to bring the best of both worlds, results on the effect of hybrid decision-makers on decision subjects' fairness perceptions are to date inconclusive. Wang et al. [419], for instance, compared hybrid *vs.* fully algorithmic decision-makers and found no significant effect of the hybrid configuration on fairness perceptions.

While the importance of human intervention has been repeatedly claimed, there is still little clarity on the causes that, counterintuitively, lead to a lack of effect of human intervention on decision subjects' fairness perceptions. There is, therefore, a need to characterize decision subjects' fairness perceptions towards human intervention.

1.3. Research Questions and Original Contributions

In this section, we formulate the research questions that we address in this dissertation and we summarize our original contributions. These research questions contribute to answering the overarching research question defined in section 1.1. Each research question (RQ) addresses a knowledge gap identified in section 1.2, i.e., RQ1 addresses knowledge gap 1, RQ2 addresses knowledge gap 2 and so on.

RQ1: What are the main trustworthy Artificial Intelligence principles and how are these principles operationalized so as to enable a multi-stakeholder deliberation in AI design and assessment? By answering to Research Question 1, contribute an overview of the current status of the trustworthy AI discourse and locate contestability within that discourse. Through research question 1, we also address the need to systematically operationalize trustworthy AI principles for these to be applicable by organizations developing and deploying AI systems. We also highlight the need to open up spaces for multi-stakeholder deliberation in AI design and assessment. The insights we generate when answering to research question 1 motivate the need to further look into the operationalization of *contestability* as a key ethical principle to ensure trustworthy AI systems.

We address research question 1 in chapter 2. Chapter 2 contributes to the multidisciplinary field of fairness, accountability, and transparency in AI. Specifically, we make the following contributions:

- We conduct a meta-review and design a framework for AI design and assessment that visualizes the main principles discussed in trustworthy AI standards. The geometrical arrangement of principles helps to identify potential similarities and tensions between principles.

- Following Value Sensitive Design approaches [147], we operationalize the identified main trustworthy AI principles into criteria and manifestations.

- We compile a collection of stakeholder-specific means to enable multi-stakeholder deliberation around trustworthy AI principles. The scarcity or lack of means illustrates future research opportunities in the field.

RQ2: What are decision subjects' needs for meaningful contestability in algorithmic decision-making? In research question 2, we focus on contestability as a key ethical principle for trustworthy AI. We adopt the broadest definition of contestability (see section 1.2.2); we do not limit contestability to one-to-one appeal or recourse processes where decision subjects can only contest their own individual decision. Contestability can be exercised by different stakeholders. Among the different stakeholders that can exercise contestability, we follow recent work in human-computer interaction (e.g., [12, 70]) and focus on individuals that are likely to be impacted by algorithmic decisions, i.e., decision subjects. Through research question 2, we generate empirical insights into decision subjects' needs for meaningfully contesting algorithmic decisions in high-risk contexts. These needs include both procedural and information needs.

We address research question 2 in chapter 3. Chapter 3 contributes to the field of human-computer interaction; specifically, to the subfield of computer-supported cooperative work. Since research question 2 does not limit contestability to one-to-one appeal and recourse processes, the cooperative aspects of contestability emerge. We make the following original contributions:

- We identify decision subjects' information and procedural needs for contestability, along with factors that might affect these needs. Information needs include *actionable* explanations with varying levels of detail. Procedural needs include mechanisms to effectively engage in contestation processes, and support structures that enable decision subjects to deal with organizational constraints. Factors that affect decision subjects' needs for contestability include AI literacy and experience with AI fairness.
- We provide a visual representation of a generic development, decision and contestation process in algorithmic decision-making. We locate the identified needs and factors in that visual representation. Such visual representation highlights the multi-stakeholder nature of contestability at a system level. It also encourages future research into contestability to account for different stakeholders' needs and interests.

RQ3: How do elements related to "the right to contest" automated decisions as interpreted from the European Union's General Data Protection Regulation (i.e., explanations, human intervention, appeal mechanisms) affect decision subjects' fairness perceptions? With research question 3, we complement the discourse on contestability in the field of human-computer interaction with interpretations provided by legal scholars on recent policy efforts. We, therefore, operationalize contestability as interpreted from Article 22(3) of the European Union's General Data Protection Regulation [131]. This involves operationalizing contestability as the *right to contest* automated decisions and related factors (i.e., explanations, human intervention). Through research question 3, we generate empirical insights into decision subjects' fairness perceptions towards different contestable algorithmic decision-making configurations.

We address research question 3 in chapter 4. Chapter 4 contributes to the field of human-computer interaction. We make the following original contributions:

- We identify explanations that combine counterfactual scenarios and the influence of input features to be the most effective type of explanations in ensuring that the information provided to decision subjects is understandable, actionable, and that it supports contestability.
- We show that the presence of explanations positively contribute to informational fairness perceptions, and that the presence of appeal mechanisms positively contribute to procedural fairness perceptions.
- We show that informational and procedural fairness perceptions predict overall fairness perceptions and demonstrate the utility of adopting a multi-dimensional approach to capturing fairness perceptions in algorithmic decision-making.
- We publish a dataset of crowdsourced perceptions of informational, procedural, and overall fairness towards algorithmic decision-making configurations with the presence of (or lack thereof) explanations, human oversight, and appeal mechanisms. The scientific community can use this data set to further investigate perceptions towards each of the elements that compose such fairness perceptions in this context and inform efforts of data collection in similar studies.

RQ4: How do varying levels of human intervention affect decision subjects' fairness perceptions in algorithmic decision-making? With research question 4, we further look into human intervention as a key factor that conditions the existence of the right to contest automated decisions as interpreted from Article 22(3) of the General Data Protection Regulation [131]. Since Article 22(3) only applies to fully automated decisions, the presence of human intervention affects the application of Article 22(3) and, therefore, the existence of a right to contest the automated decision by decision subjects. Effective human intervention is necessary to design algorithmic decision-making processes that attend decision subjects' rights, freedoms, and interests in the absence of a right to contest automated decisions. Through research question 4, we generate empirical insights into decision subjects' fairness perceptions towards algorithmic decision-making configurations with varying levels of human intervention. To this end, we first explore appropriate models to capture decision subjects' perceptions towards decision-makers, and evaluate how these perceptions relate to fairness perceptions.

We address research question 4 in chapter 5. Chapter 5 contributes to the field of human-computer interaction. We make the following original contributions:

- We identify decision-maker profile, model type, and data provenance as prominent decision-maker-related characteristics that decision subjects consider when developing perceptions towards decision-makers (study 1).
- We identify the Ability, Benevolence, and Integrity model as a useful methodological approach for capturing perceptions towards decision-makers (study 1).
- We generate empirical insights that show that (1) the decision-maker's profile affects perceived ability, benevolence, and integrity; (2) perceived ability, and integrity positively relate to decision subjects' fairness perceptions; (3) the effect of

the decision-maker's profile on decision subjects' fairness perceptions is mediated by perceived ability and integrity (study 2).

- We publish a dataset of crowdsourced perceptions of ability, benevolence, integrity, and fairness towards different decision-making configurations. The scientific community can use this data set to further investigate perceptions towards algorithmic decision-making in this context and inform efforts of data collection in similar studies.

1.4. Research Approach

In this section, we outline the research approach taken to conduct the research we present in this dissertation. We discuss (1) how we capture decision subjects' needs and perceptions to inform the development and deployment of contestable AI, (2) the mixed-methods approach we adopt, (3) how we ground the conducted studies in decision-making contexts both in the public and private sectors, as well as (4) how we follow open science principles.

1.4.1. Capturing Decision Subjects' Needs and Fairness Perceptions to Inform the Development and Deployment of Contestable AI

In this dissertation, we capture decision subjects' needs for and fairness perceptions towards contestability as an approach to create knowledge for informing the development and deployment of contestable AI. Our approach is part of a broader effort to ensure a human-centered development and deployment of AI systems [374]. Recent human-centered studies account for decision subjects' needs and perceptions in AI design processes, and have become popular in the field of human-computer interaction (e.g., [70, 407]). The objective of considering decision subjects' needs and perceptions in the design of AI (either by directly involving these decision subjects in participatory approaches or by indirectly capturing those needs and perceptions through user studies) is to better support their interests and goals from the very early stages of the design process [251]. By accounting for decision subjects' interests and goals, in turn, we aim to avoid many of the harmful and unwanted effects that AI systems might produce [377]. While this might slow down the deployment of AI systems, it should be successful in ensuring a fair and effective development of such systems.

Two distinct stages can be identified in our approach: a generative stage and an evaluative stage. We use the term *generative* to refer to the stage where we do not limit contestable AI to any existing approach. We generate ideas for the design of contestable AI systems by addressing decision-subjects' needs and desires. We use the term *evaluative* to refer to the stage where we assess solutions for contestable AI (by capturing decision-subjects' fairness perceptions) suggested in policy efforts such as the GDPR [131]. Chapter 3 is part of the generative stage. In chapter 3, we run interviews with decision subjects and capture their needs for contestability. While we prompt participants with an existing AI system, we do not limit the contestation process to any pre-existing workflow. The objective of this generative stage is to learn from our participants' lived experiences and map decision subjects' most prominent needs for contestability throughout the AI lifecycle.

Instead, chapters 4 and 5 are part of an evaluative stage. In chapters 4 and 5 we interpret the right to contest automated decisions and the safeguard of human intervention from Article 22(3) of the General Data Protection Regulation [131]. The objective of this stage is to evaluate the adequacy of contestation mechanisms suggested by current policy efforts. To this end, we draw from literature in organizational psychology and we capture decision subjects' fairness perceptions towards different algorithmic decision-making configurations. Literature in organizational psychology has shown that capturing employees' fairness perceptions is an effective way of predicting employees' attitudes and behaviors [96]. Employees are more willing to accept negative decisions if they perceive that the process and treatment leading to a decision are fair [95]. For policy enforcement, it has also been shown that fairness perceptions positively relate to perceived legitimacy and long-term compliance with decisions that legal authorities make [399]. In organizational psychology, fairness perceptions are captured along up to four justice dimensions. *Distributive justice* deals with the equity in decision outcomes distribution. *Procedural justice* deals with the processes that lead to those outcomes. *Informational justice* deals with the information provided to decision subjects about the decision-making process. *Interpersonal justice* deals with the treatment received by decision subjects as part of the decision-making process. Capturing decision subjects' fairness perceptions represents a first step towards crafting algorithmic decision-making processes that represent legitimate and acceptable alternatives to human-led decision-making.

1.4.2. A Mixed-Methods Approach

This dissertation relies on a combination of *qualitative* and *quantitative approaches* to generate empirical insights into decision subjects' needs for contestable Artificial Intelligence systems and to capture decision subjects' fairness perceptions. A mixed-methods approach allows us to build a comprehensive understanding of decision subjects' perspectives on contestability. On the one hand, thanks to the open-ended nature of qualitative approaches, we generate nuanced insights that help us decide the way in which we should move forward with the interpretation of the phenomenon at hand, generate new research questions and formulate hypotheses [173]. On the other other hand, quantitative approaches enable us to narrow down the field of study, and test the research questions and hypotheses formulated through qualitative studies.

Qualitative approach. In chapter 3 and chapter 5 (study 1) we conduct qualitative interviews prompted by vignettes. The aim of conducting these interviews is (1) to capture decision subjects' needs for contestability and (2) to identify characteristics of decision-makers that might affect decision subjects' fairness perceptions, respectively. Qualitative interviews are appropriate means for capturing participants' needs and perceptions when participants have something at stake in the selected context [90]. Vignettes are fictitious descriptions of events related to the topic of study [43, 351]. Using vignettes for capturing participants' reactions in a particular context is widely accepted in social research [43, 192].

In both chapter 3 and chapter 5, we explore a context in the public sector where AI systems are used for the identification of illegal holiday rentals. Participants are individuals with experience renting their homes out for short-term stays. They are located in municipalities in Western countries where there has been considerable effort in setting up methods and workflows for detecting illegal holiday rentals. The vignettes we use as part of the interviews describe a scenario where the municipality (through AI systems) find participants' property to be illegally rented as a holiday rental. This allows us (1) to capture participants' reactions and perceptions to the use of AI for identifying illegal holiday rentals and (2) to identify their needs for meaningful contestability as impacted individuals, i.e, decision subjects.

We analyze the conducted qualitative interviews through *reflexive thematic analysis* [68, 91]; by combining an inductive and deductive orientation to data. Thematic analysis is a method that aims at answering a research question by identifying patterns and themes in a dataset [90]. Reflexive approaches to thematic analysis acknowledge the considerable "analytic and interpretative work" [67] of the researcher(s) when identifying patterns of shared meaning (i.e., themes). Reflexive thematic analysis involves six main steps: (1) transcription of recordings, (2) familiarization with the data, (3) grouping quotes in codes and code groups, (4) searching for themes, (5) crafting and mapping themes, (6) refining codes based on those themes [91]. Reflexive thematic analysis represents a flexible method that allows for an in-depth engagement with the data. Compared to other forms of thematic analysis (e.g., codebook [221], coding reliability [65]), *reflexive* thematic analysis is the most suitable approach when researchers want to explore "deep, complex, nuanced meaning and understanding" *in data* [91]. Since in chapters 3 and chapter 5 (study 1) we aim at identifying patterns *in data* and interpreting them [67], reflexive thematic analysis is an appropriate approach to use.

Positionality Statement. Reflexivity acknowledges the role of researcher(s) when analyzing, interpreting, and making sense of qualitative data [90]. As a researcher coming from a Southwestern European country, currently living in a Northwestern European country, and working at a public technical university, I acknowledge that my perspectives and lived experiences shape the knowledge I generate through qualitative research. My disciplinary background in industrial (mechanical) engineering also shapes the qualitative research I conduct. The backgrounds and lived experiences of my co-authors, also partially involved in data analysis, equally shape the nature of the performed analysis.

Quantitative approach. In chapter 4 and chapter 5 (study 2) we conduct quantitative crowdsourced studies prompted by vignettes. The aim of the quantitative crowdsourced study in chapter 4 is to (1) identify explanations that are understandable, actionable and that support cotestability, and (2) to evaluate the effect of explanations, human oversight, and appeal mechanisms on decision subjects' fairness perceptions. We follow literature in organizational psychology, and adopt a multi-dimensional approach to capturing decision subjects' fairness perceptions [95]. We specifically focus on capturing the effect of explanations, human intervention (in the form of human oversight) and appeal mechanisms on informational and procedural fairness perceptions.

The aim of the quantitative crowdsourced study in chapter 5 (study 2) is to (1) capture decision subjects' perceptions of ability, benevolence, and integrity towards different decision-making configurations with varying levels of human intervention and, to

(2) evaluate the relation between those perceptions and decision subjects' fairness perceptions. Quantitative crowdsourced studies allow us to capture the subjective perceptions and understandings of a big pool of participants.

In chapter 4, we explore a context in the private sector where AI systems are used for driving or augmenting a loan approval workflow. In chapter 5 (study 2), we explore a context in the public sector where AI systems are used for identifying illegal holiday rentals. In both cases, participants are individuals who (1) are at least 18 years old, (2) are proficient in English, (3) are located in the Global North, and (4) participate in the study only once. Participants who do not pass the corresponding attention checks are excluded from analysis. Participants in our studies are incentivized through monetary rewards. Vignettes, in these two quantitative studies, act as catalysts for participants to adopt the role of decision subjects and for subjective perceptions towards the decision-making process to manifest in relation to each specific context.

We use statistical significance tests to either accept or reject the formulated hypotheses. We use both parametric and non-parametric tests. We use parametric tests if the assumptions of (1) normality, and (2) homoscedasticity (i.e., homogeneity of variances) are met. If these assumtion are not met, we use non-parametric tests instead. While mainly quantitative, open-ended questions that complement the conducted studies are analyzed using *reflexive thematic analysis* [68, 91].

1.4.3. Grounding Research in Public and Private Decision-Making Contexts

In this dissertation, empirical work is grounded in two different contexts; the first one in the public sector and the second one in the private sector. Grounding the research presented in this dissertation in both contexts will shed some light on the representativeness of each case. Similarly, it will help determine the way in which the results we get in one of the contexts are transferable to the other context and vice versa.

1. **Public sector: illegal holiday rental identification.** The use case in the public sector deals with an Artificial Intelligence system (and corresponding workflow) suggested by the municipality of Amsterdam to detect illegal holiday rentals. Even if the municipality decided not to deploy this system, it was originally suggested as a means for augmenting civil servants' capacities to investigate potential illegal holiday rentals after receiving a report on a particular address. The system, which is based on a random forest model, is designed to compute the probability that the reported property has of being an illegal holiday rental. To this end, the system relies on data about (1) the identity of the owner, (2) the building, and (3) previous illegal housing cases. The computed probability is considered by civil servants as a relevant factor when deciding whether to further investigate the report.

2. **Private sector: loan approval scenario.** The use case in the private sector follows prior work [53, 59, 271, 304, 365] and deals with a generic AI system used as part of a loan approval scenario. The system relies on relevant factors such as annual income, credit score, employment status or the requested amount to evaluate decision subjects' eligibility for a loan. The decision to grant or not to grant the loan is

presented to the decision subject along with information (if explanations are provided) about the factors that helped the financial company to make that decision and the weight that each factor had on the final decision.

When dealing with decision-making processes, the public and private sectors are distinct. Three main differences are worth highlighting. First, if decision subjects get their loan request rejected in a financial company, they have the alternative to go to another company and get their eligibility for a loan re-evaluated. In the case of the public sector, instead, citizens lack any alternative to dealing with administration [11]. Second, unlike processes in the private sector, administrative processes in the public sector require civil servants to retain a level of autonomy that allows them to evaluate decision subjects' individual circumstances as part of the decision-making process, i.e., they apply discretion [356]. Third, decision-makers in the public sector need to make societally sensitive decisions while operating under bureaucratic constraints in a multi-actor playing field. This is not necessarily the case in decision-making processes in the private sector.

1.4.4. Open Science

All empirical studies included in this dissertation have been pre-registered. The motivation behind preregistering these is (1) to uphold transparency throughout the research process, and (2) to enable the conducted studies to be reproduced and/or scrutinized by the scientific community [79]. The utility of preregistering qualitative studies is different from the utility of preregistering quantitative studies.

- In this dissertation, we conduct **qualitative studies** to better understand the phenomenon at hand and to generate hypotheses based on the collected data (i.e., postdiction research). To this end, we engage in a cyclic process of collecting and analyzing the data. We acknowledge the subjectivity inherent to qualitative research and our role as researchers in making sense and interpreting the data. We also acknowledge the importance of a flexible approach to qualitative research [173]. The utility of preregistering the qualitative studies in this dissertation lies in (a) describing the original aims of the study based on theory that is relevant to the topic, (b) registering the presuppositions that underlie the data collection and analysis processes, and (c) enabling the scientific community to track the development of the study [173].

- In this dissertation, we conduct **quantitative studies** to test hypotheses on the collected data (i.e., prediction research). Hypotheses are generated based on theory, previous work, and the exploratory insights generated through qualitative studies. The utility of preregistering the quantitative studies in this dissertation lies in (a) "freezing" the formulated hypotheses at time t_0 before data collection so that postdictions are not presented as predictions, (b) registering the research questions and analysis plan before data collection, (c) allowing the scientific community to scrutinize the study design and plan [173].

For the sake of transparency, reproducibility, and rigor, the instruments used to conduct all empirical studies and the generated (anonymized) data are also available for the broader scientific community through the Open Science Framework (OSF). Chapters 3 to 5 include the corresponding links to OSF.

1.5. Thesis Structure and Chapter Origins

Chapter 2 presents a meta-review of principles for trustworthy Artificial Intelligence. These principles are operationalized by breaking them down into criteria and manifestations. This meta-review leads to a framework covering prominent ethical principles, and encouraging multi-stakeholder deliberation to deal with tensions inherent in the design and assessment of socio-technical systems. Chapter 2 is based on the following conference paper:

- Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. "Towards a multi-stakeholder value-based assessment framework for algorithmic systems". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* (FAccT'22)

Chapter 3 focuses on contestability as a prominent ethical principle for trustworthy AI. It generates empirical insights into decision subjects' information and procedural needs for meaningful contestability. Chapter 3 is based on the following research paper:

- Mireia Yurrita, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, Alessandro Bozzon. "Personalize, Prioritize, Collectivize: Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability". In *Proceedings of the ACM on Human-Computer Interaction CSCW* (CSCW '25).

Chapter 4 captures decision subjects' fairness perceptions towards different contestable algorithmic decision-making processes. Chapter 4 narrows contestability down into the *right to contest* automated decisions as interpreted from Article 22(3) of the General Data Protection Regulation (GDPR) [131]. This involves algorithmic decision-making configurations with the presence (or lack thereof) explanations, human intervention, and appeal mechanisms. Chapter 4 is based on the following conference paper:

- Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. "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* (CHI'23).

Chapter 5 further focuses on human intervention as a safeguard defined in Article 22(3) of the General Data Protection Regulation (GDPR) [131] that conditions decision subjects right to contest automated decisions. Chapter 5 first identifies appropriate models to capture perceptions towards decision-makers in algorithmic decision-making. It then captures decision subjects' fairness perceptions towards algorithmic decision-making configurations with varying levels of human intervention. Chapter 5 is based on the following conference paper:

- Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia Pont, Alessandro Bozzon. "Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI'25).

Chapter 6 summarizes the results of this dissertation, discusses the broader implications of this work and reflects on its limitations.

2

Operationalization of Trustworthy AI Principles

In this chapter, we contribute an overview of the main trustworthy Artificial Intelligence (AI) principles and we operationalize these principles so as to enable multi-stakeholder deliberation for AI design and assessment (**RQ1**). We develop a framework that covers prominent ethical principles for trustworthy AI. This framework presents a circular arrangement of ethical principles with two bipolar dimensions that make common motivations and potential tensions explicit. In order to operationalize high-level principles, these are broken down into specific criteria and their manifestations. However, some of the criteria are mutually exclusive and require negotiation. Instead of merely relying on AI researchers' and practitioners' input, we argue that it is necessary to include stakeholders that present diverse standpoints to systematically negotiate and consolidate tensions across principles and related criteria. To this end, we map stakeholders with different insight needs, and assign tailored means for communicating manifestations of trustworthy AI principles to them. We, therefore, contribute to current trustworthy AI discourse with an assessment framework that visualizes closeness and tensions between ethical principles and we give guidelines on how to operationalize them, while opening up the design, evaluation and deliberation process to a wide range of stakeholders.

This chapter is based on the following publication: **Mireia Yurrita**, Dave Murray-Rust, Agathe Balayn, Alessandro Bozzon. "Towards a multi-stakeholder value-based assessment framework for algorithmic systems". In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22).

2.1. Introduction

In recent years, it has become clear that Artificial Intelligence (AI) systems might encode harmful biases and might lead to unfair outcomes [298, 372]. The dangers of using Machine Learning (ML), specifically, in Computer Vision (CV) [74] or Natural Language Processing (NLP) [16, 49, 106, 240, 353], for assessing recidivism [396], for candidate screening [332] and for recommending content on social media platforms [208, 320, 337, 446] have been pinpointed. The origins of harmful algorithmic bias[1] might be diverse [372, 387]. Just to mention a few, representativeness issues, play a key role in disparate algorithmic performance [4, 83]. The way in which data is collected [49, 325] and labelled [99, 118, 325] is a major menace to data soundness. Beyond the data generation process, aggregation, learning, evaluation and deployment biases have been identified throughout the ML pipeline [387]. In response to harmful algorithmic bias, current auditing processes [2] [3, 349, 425] have provided numerous useful bias detection techniques [18, 37, 47, 124, 155, 191, 430, 432, 439].

However, harmful algorithmic behavior is not limited to biases encoded in the AI life cycle [372]. The lack of social and cultural context in the mathematical representation of socio-technical systems [278, 372] or the omission of changing practices and long-term effects of the deployed systems [64, 103, 196, 234] are also some problematic aspects that are hardly considered in current design and auditing processes. Such processes mostly mostly assess the conformance of AI systems to applicable standards [195] through quantitative analysis, rather than additionally gaining insights into their contextual implications [334, 372]. Furthermore, these design and auditing approaches solely rely on AI researchers, and practitioners, who can fail to detect issues that arise from context-dependent unanticipated circumstances during usage time [372].

In this chapter, we argue that: firstly, design and assessment processes for AI systems should go beyond bias evaluation and take into account additional high-level values [3] that are outlined in Artificial Intelligence (AI) ethics documents [20, 85, 132, 138, 159, 194, 201, 289, 315, 391, 402]. Contestability, for example, has been identified as a key value of algorithmic systems, but there is still little guidance on what contestability requires [269]. In order to provide a good coverage of values that deal with principled algorithmic behavior, we develop a value-based design and assessment framework, where contextual conditions are considered along with quantifiable measurements. We organize such values in a circular layout with two bipolar dimensions. As claimed by Friedman et al. [147], values do not exist in isolation. They are situated in a delicate balance and touching one value might have implications in another value [147]. This means that

[1] Following the approach adopted by Shen et al. [372], we will distinguish between harmful algorithmic biases and harmful algorithmic behaviors, since not all harmful algorithmic behaviors originate from biases and not all algorithmic biases are necessarily harmful [63].

[2] We will use the term *auditing* processes to refer to external audits, where third parties only have access to model outputs [352]. We will use the term *assessment* processes to refer to an evaluation process that is applied "throughout the development process and that enables proactive ethical intervention methods" [334]. We will not use the term *Internal Audit* defined by Raji and Smart [334] to avoid erroneous inferences that would limit the stakeholders of our framework to the employees of an organization.

[3] We will adopt the definition of *values* used in philosophy of science, following Birhane et al. [61]. Values of an entity are, thus, defined as properties that are desirable for that kind of entity. In this chapter, we will use the term *values* to refer to the ethical principles that aim at ensuring AI systems are trustworthy.

value interactions need to be taken into account when making choices about value prioritization and situating algorithmic systems in a space of trade-offs [30]. The circularity of our framework makes such interactions explicit and facilitates the identification of common motivations and tensions among values.

Secondly, AI design and assessment processes should give tangible guidelines for the operationalization [4] of values, so as to eventually put ethics into practice following a context-aware approach [373]. To this end, each value in our framework is broken down into criteria manifested through quantifiable indicators, process-oriented practices or signifiers[5]. These value-specific criteria and their manifestations can be used either as a checklist if our framework is applied for evaluating a system that is already developed, or for promoting such values if it is being used during design time.

Thirdly, AI design and assessment processes should allow critical reflection on AI systems and engage in conflictual plurality[6]. Inevitable value tensions inherent in the nature of socio-technical systems [167] require spaces for ethical discussions [373], that can benefit from the insights of multiple stakeholders beyond AI practitioners [30, 372]. To enable fruitful multi-stakeholder discussions [245], we map and match value-specific communication means with different stakeholders. We, therefore, contribute with:

- A review of prominent high-level values in AI ethics and translation into specific criteria through the:
  - Development of a design and assessment framework that facilitates the identification of common motivations and tensions among values encoded in AI systems.
  - Definition of guidelines to deal with the complex middle ground between abstract values and concrete system specifications.
- Translation of value-specific criteria into manifestations that are understandable for diverse stakeholders through the:
  - Review of available means to communicate value manifestations to different stakeholders based on their insight needs and nature of knowledge.
  - Definition of steps to introduce those communication means into multi- stakeholder deliberation dynamics.

The remainder of the chapter is organized as follows: in section 2.2, we analyze related work for documenting and auditing AI systems. We also introduce the theoretical basis of our framework. Section 2.3 describes and justifies the selected values, criteria and manifestations and their arrangement in our framework. Section 2.4 maps the

[4] Our strategy follows the definition by Shahin et al. [369], where "operationalizing values" is defined as the process of identifying values and translating them into concrete system specifications that can be implemented.

[5] We adopt the definition given by Don Norman in his 2013 edition of "The Design of Everyday Things". Signifiers are perceivable cues of an affordance, affordances being "the relationship between the properties of an object and the capabilities of the agent that determine how the object could be possibly used". In this chapter, the "object" in question is the AI system.

[6] We understand *conflictuality* as a solution for dealing with the "figure of alterity". Unlike *conflict*, it represents a method for linking opposing views and opening out onto the exercise of thinking [149]

stakeholders involved in the AI design and evaluation process, and reviews the available means for communicating system-specific information to them. In sections 2.3 and 2.4, we illustrate the necessary steps for navigating our framework through an example in the area of life insurance application. We discuss our approach, its implications, and future lines of work in section 2.5, and we conclude this chapter in section 2.6.

2.2. Background and related work

In this section, we survey current practices for documenting and auditing technical specifications of AI systems. We also provide the theoretical basis of our framework.

2.2.1. Background

Standardized documentation.

In order to facilitate the audit of AI systems, it is important that technical specifications are collected and documented in a standardized way. So far, AI system documentation practices are limited to datasets and models.

Documenting datasets. Recent studies in documentation practices for AI datasets claim the need for greater data transparency [193]. Since the quality of the prediction made by the AI system highly depends on the way the data has been collected, the need for setting rigorous practices (as it is the case in other areas of knowledge, such as social sciences or humanities [153]) has been highlighted [325]. Likewise, the choice of what data to collect and how to collect this data is in itself a value-laden decision [110, 360]. To standardize documentation for AI datasets and make data-related decisions more transparent for other practitioners, various methodologies have been suggested in the last years, "Datasheets" [152] and "Dataset Nutrition Labels" [187], for instance. For NLP techniques, "Data Statements" are regarded as a dataset characterization approach that helps developers anticipate biases in language technology and understand how these can be better deployed [50].

Documenting models. In addition to documenting datasets, the importance of disclosing the technical characteristics of AI models has also been emphasized. A good example of model documentation practices are the "Model Cards" [294].

Auditing techniques.

Various methodologies and tools for incorporating auditing tasks into the AI workflow have been suggested. Aequitas [349] is an open source toolkit to detect traces of bias in models. The toolkit designed by Saleiro et al. [349] facilitates the creation of equitable algorithmic decision-making systems where data scientists and policymakers can easily use Aequitas for model selection, evaluation and approval. Wilson et al. [425] described a framework that helps ensure fairness in socio-technical systems, and used it for auditing the model of the startup *pymetrics*. Adler et al. [3] studied auditing techniques for black-box models to discover whether proxy variables linked to sensitive attributes indirectly influence the predictions of the system. The end-to-end "Internal Audit Framework" suggested by Raji and Smart [334] is of special interest for justifying the need of setting specific guidelines to enable multi-stakeholder deliberation in design and assessment processes. It consists of five main stages where the need for stakeholder diversity is highlighted, e.g. the scoping stage calls for covering a “critical range of viewpoints” to review the ethical implications of the system use case.

Motivation.

While standardized documentation practices [50, 152, 187, 294] and audits [3, 349, 425] have been influential methodologies for dealing with harmful algorithmic bias, their scope is limited to performing quantitative analysis over data and model outputs so as to ensure compliance with applicable standards [195]. Such an approach does not deal with additional ethical values which cannot be easily quantified [245] and that are essential for ensuring desirable algorithmic behavior. One could argue that “Datasheets” [152] and “ModelCards” [294] already devote a section to the description of ethical considerations of datasets and models. Yet, there are no specific guidelines on how to identify ethical issues. As Shklovski et al. [373] discovered, technical people both in industry and academia struggle to identify what an ethical issue entails. To address this caveat, as part of our value-based framework, we give tangible guidelines for putting ethics into practice [298, 373]. We operationalize each high-level value into actionable value criteria and their manifestations. One could also argue that Raji and Smart [334] already included an Ethics Review as part of their end-to-end internal audit framework. Indeed, they exemplified such a review by describing ethical considerations and potential mitigation strategies against bias and privacy threats for a smile detection system. However, this review does not address most of the values that are referred in AI ethics documents. We fill in this gap by offering a good coverage of values to examine, including those that normally go unnoticed in current documenting and auditing practices.

2.2.2. Accounting for human values in the design and assessment of AI systems

Our design and assessment framework identifies and arranges values encoded in AI systems by covering prominent principles in AI ethics and organizing them in a circular structure.

Addressing human values in technology.

For the definition of our value-based framework, we followed other theoretically grounded approaches, such as Value Sensitive Design (VSD) [147]. VSD represents a pioneering endeavour where human values are proactively considered throughout the process of technology design [107]. Just as VSD does with interactive systems, we address the need to account for human values during the design, implementation, use, and evaluation [107] of AI systems. To this end, we select and define values involved in AI systems, and we identify stakeholders that will be in contact with such systems and whose standpoints need to be considered. Our approach resonates with conceptual investigations described in VSD literature [107].

The circular nature of our framework is inspired by Schwartz’s Theory of Basic Human Values [366]. This theory identifies individual value priorities based on ten basic personal values. Values are arranged in a circular form and categorized in four quadrants. These quadrants are located in two bipolar dimensions, which visualize “oppositions between competing values”. In addition, adjacency between values denotes a common motivation, which results in these values forming a circular continuum. The advantage of adopting a circular arrangement, like the one suggested by Schwartz, for AI systems is that value commonalities and trade-offs can be easily identified thanks to their positioning. Considering the struggles of technical people when addressing ethical issues [373], an explicit representation of value interactions will facilitate the analysis of trade-offs and decision-making about value prioritization.

Ethical principles for AI systems

The values considered in our AI design and assessment framework cover prominent principles outlined in AI ethics. In the last five years, many institutions have studied and defined high-level principles that AI systems should follow [138]. As a matter of fact, documents that aim at guiding the "ethical development, deployment and governance of AI" are converging into a common set of principles [296, 298]. However, high-level principles are far from being actionable [298] and it is necessary to provide answers on how to proceed [5]. Efforts for going from "what" to "how" [7] include the review carried out by Morley et al. [298], where available tools for operationalizing ethical principles were examined. Similarly, the AI Ethics Impact (AIEI) Group designed a framework for rating the presence of ethical principles in AI systems, getting inspiration from energy efficiency labels [4].

Our value-based framework differs from previous applied ethics frameworks [4, 298] in various ways. Firstly, we arrange values in a circular form, which makes it easier to navigate common motivations and trade-offs between values. Although such common motivations and trade-offs can be inferred from current AI ethics documents, we make them explicit by arranging values in a geometrically meaningful way. This is especially useful for identifying overlaps between values that are adjacent to each other and for detecting potential value tensions that need to be negotiated and consolidated. Secondly, we do not limit our ethics framework to a mere checklist. We follow Shklovski et al. [373] and combine the enumeration of tangible and actionable value manifestations with the generation of an open space for ethical debate. As opposed to the deterministic approach adopted by the AIEI group [4], we map communication means for facilitating ethical reflections of AI systems and for addressing ethical issues in practice [373]. Thirdly, as opposed to previous applied ethics frameworks [4, 298], we embrace diversity in ethical reflections and deal with the complexities that arise from plurality. In order to facilitate multi-stakeholder discussions, we match available communication means for addressing different value manifestations with stakeholders that present different insight needs.

2.3. Design of our value-based framework

In this section, we describe the composition of our value-based framework and justify its arrangement. We provide the definition of each of the selected values and the derived criteria and manifestations.

[7] Expression used by Morley et al. [298] to refer to the operationalization of ethical principles in AI. The 'what' refers to the ethical principles themselves, whereas the 'how' refers to the act of putting such principles into practice.

2.3.1. Methodology for reviewing values, criteria and manifestations in AI systems

To design our framework, we analysed documents outlining high-level ethical principles that AI systems should follow. Our starting point was the review performed by Fjeld et al. [138], where principles coming from governments, inter-governmental organizations, multiple stakeholders, the private sector, and the civil society were examined. In their review, Fjeld et al. identify nine key themes, some of which overlap with the values outlined in our framework. The identification of prominent high-level values was also complemented with other reviews [4, 61, 105, 181, 298]. To identify the criteria that define the fulfilment of prominent high-level values, we navigated the visual representation provided by Fjeld et al. [138] and accessed the documents that offer a higher coverage of the value in question. For instance, for the value of *privacy*, one of our main references has been the GDPR [131].

We went from criteria to value manifestations through an extensive exploration of available value-specific reviews that identify such manifestations. For instance, for the value of *security* Xiong et al. [429] presented a thorough study of mechanisms used for securing the ML pipeline against external threats. For *explainability*, Barredo-Arrieta et al. [42] put together more than four hundred references and mapped strategies in the field of Explainable Artificial Intelligence [42]. We partly rely on such reviews for identifying value manifestations because our contribution lies in covering and putting together a set of values and their manifestations in AI systems to end up with a "health-check" for assessing AI systems, rather than rediscovering such value manifestations ourselves. Similarly, for the values of *performance* and *fairness*, we only included the main value manifestations that represent the basis for any other derived metrics. That is to say, just as Verma et al. [414] did, we outline the main quantifiable indicators (false positives, false negatives etc) used for measuring *performance* and *fairness*, but we are aware that many other metrics that derive from these ones can be insightful for specific contexts. Dealing with such compound metrics is out of the scope of this work.

2.3.2. Design and assessment of AI systems through a circular value-based framework

Our resulting AI design and assessment framework arranges values in a circular form (figure 2.1). Adjacency between values denotes a common motivation and oppositions between competing values are represented through two bipolar dimensions. For instance, adjacency between *privacy* and *security* denotes a common objective towards the protection of sensitive information [138, 347] and resilience to external threats [289]. The trade-off between *privacy* and *explainability*, on the other hand, is made explicit by their opposing positioning in our circular framework. High-level values are then broken down into specific criteria and their manifestations, as indicated in figure 2.2. Criteria defining a specific value ultimately represent a set of questions to be asked as part of the design and assessment process to ensure the fulfillment of the value in question —if the framework is being applied before deployment— or the promotion of a specific value —if the framework is being applied during design time—. These sets of criteria are not unique and exclusive to one value. For instance, when defining the criteria for *privacy* we refer to "data protection", which is also involved in *security* in the form of "resilience

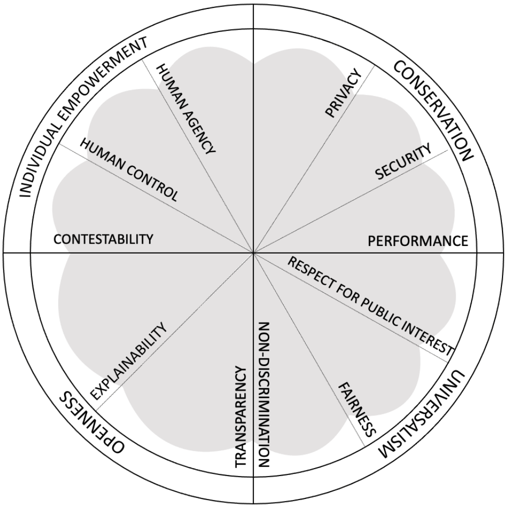


Figure 2.1: Graphic representation of our circular value-based design and assessment framework. Oppositions between competing values are illustrated through the arrangement of those values in bipolar dimensions and common motivations through adjacency between values, which form a circular continuum.

to attacks". These overlaps are precisely what we want to highlight and make explicit thanks to the circularity of our framework and adjacency between values.

Manifestations are classified in three groups depending on their nature: (1) *Quantifiable indicators* are specific measurable parameters that numerically manifest the (lack of) adequacy in the standards set for a criterion (magenta). (2) *Process-oriented practices* are actions and mechanisms implemented during the AI development or deployment process that advocate for a certain value (olive). (3) *Signifiers* [8] are files and reports that describe the relationship between the properties of the AI system and humans that determine how that system can be used (orange). There is a many-to-many relationship between criteria and manifestations. In the next subsections, we present opposing value categories in pairs.

Motivating example. For illustrative purposes, we guide the reader through each stage of our framework with a hypothetical yet plausible use case. Consider a team of researchers is developing an AI system for automating life insurance application processes. The system shall accept or refuse the request of a life insurance based on the following data: physiological information of the candidate, details about their employment, insurance history, and individual and family medical history. As part of a new wave of ethical finance companies, the team would like to ensure that their work is ethically grounded. However, it is not clear what that means in practice. From looking at prominent literature, they can develop a sense that the model should be fair and unbiased, and potentially that there should be some level of human control or intervention possible. Yet, they cannot be sure if they have a good set of representative values covered, and they

[8] Check footnote 5

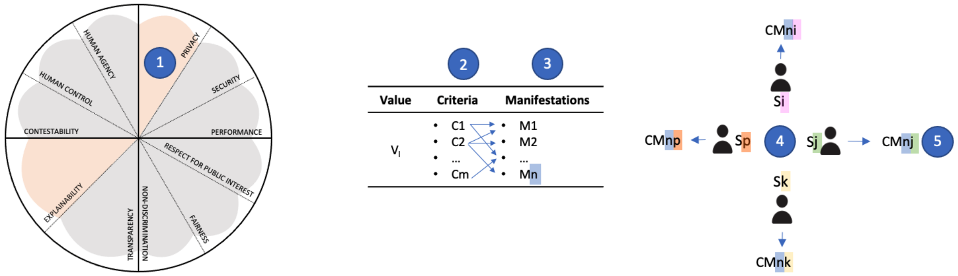


Figure 2.2: Workflow for operationalizing high-level values and for enabling multi-stakeholder design and assessment of AI systems. This workflow represents the methodology that we followed for structuring our framework and the steps that researchers and practitioners should take to make use of it. (1) Select and discuss project-specific values (V), (2) Decide on criteria (C) for embodying those values, (3) Select the manifestations (M) that enact value-specific criteria, (4) Map relevant stakeholders (S) to enable ethical reflection of value and criteria tensions, (5) Match adequate communication means (CM) to stakeholders.

do not know how to go about communicating the way that their model embodies those values. They are now reading through our multi-stakeholder value-based design and assessment framework.

2.3.3. Conservation vs Openness

The first dimension of our value-based framework captures the conflict between *conservation* and *openness*. Values included within the *conservation* category emphasize the necessity of AI systems to preserve confidentiality with regards to information, as well as, the need for the system to preserve adequate robustness when it comes to performance. On the contrary, the category of *openness* encompasses values that advocate for making system components and specifications more accessible to the public.

Conservation.
Privacy, security and performance uphold confidentiality and robustness within AI systems [138].

Privacy. The defining goal of *privacy* is the need for AI systems to respect individual's informational confidentiality [4, 138] as part of their user rights [61]. When applying this value to the AI development pipeline, data processing itself should integrate privacy standards [4, 138, 298], so that there is no possibility of identifying sensitive information about individuals [181, 415]. Furthermore, the need to provide humans with agency over their data is emphasized [138]. Based on these definitions, we identified six main criteria for the fulfillment of *privacy* within AI systems (table 2.1). (1) Consent for data usage [4, 131, 138]: data subjects should be appropriately informed when their data is being used and their explicit approval is needed. (2) Implementation of data protection mechanisms [4, 138, 141]: during the development of AI systems, resources should be devoted to making user data management secure and confidential. (3) Users having control over their data and ability to restrict its processing [131, 138]: users should be able to limit the way their personal data is being used. (4) Users having the right to rectify [4, 131,

Value	Criteria	Manifestations
Privacy	• **Consent for data usage** [4, 131, 138] • **Data protection** [4, 138, 141] • **Control over data / ability to restrict processing** [131, 138] • **Right to rectification** [4, 131, 138] • **Right to erase the data** [4, 131, 138] • **Right of access by data subject, data agency** [131, 390]	• Written declaration of consent [131] • Description of what data is collected [283] • Description of how data is handled [283] • Purpose statement of data collection [283] • Statement of how long the data is kept [283] • Form and submission mechanisms to object data collection and to make complaints [62] • Obfuscation of data [4]

Table 2.1: Illustration of how to move from values, to criteria and their manifestations with an example for *privacy*.

138]: users should be able to modify their data at any time. (5) Users having the right to erase their data [4, 131, 138]: this criterion refers to the right that users have to be forgotten. (6) Users having right to access their data [131, 390]: this right empowers users to have agency over their data. These criteria manifest in various ways. Signifiers include: a written declaration of consent [131], detailed descriptions of the collected data, how data is handled, how long it will be kept and the purpose of collecting that data [283]. These signifiers are necessary for users to fully understand what sharing their data entails. Process-oriented practices include the obfuscation of data [4] and forms and submission mechanisms to object data collection and make complaints [62].

Security. Definitions characterizing *security* (see table 2.2) highlight the need for AI systems to be (1) resilient to potential maleficent attacks [138, 298] and to present a (2) predictable [4, 132, 138] and (3) robust [4] behavior at any time. This includes implementing mechanisms to protect user privacy, such as strategies that ensure that inferences about an individual cannot be made by interrogating the model [181, 289, 415]. Following the survey performed by Xiong et al. [429], different methodologies that aim at protecting AI systems against external threats (process-oriented practices) have been classified into two main groups. The first group consists of defence methods against integrity threats at two different stages of the AI pipeline: during training time [58, 100, 157] and during prediction time [58, 158, 272, 319]. The second group aim at defending the AI system against privacy threats, namely membership inference attacks [122, 204, 305, 375, 435].

Value	Criteria	Manifestations
Security	1. **Resilience to attacks**: protection of privacy [181, 289, 415], vulnerabilities, fallback plans [4, 138, 159, 298] 2. **Predictability** [4, 132, 138] 3. **Robustness / reliability**: prevent manipulation [4]	AGAINST INTEGRITY THREATS [429]: • Training time [429] Ex.: – Data sanitization [9] [58, 100] – Robust learning [10] [58, 157] • Prediction time [429] – Model enhancement [58, 158, 272, 319] Ex.: ◇ Adversarial Learning [11] ◇ Gradient masking [12] ◇ Defensive Distillation [13] AGAINST PRIVACY THREATS [429]: • Mitigation techniques [305]: – Restrict prediction vector to top k classes [14] [375] – Coarsen the precision of the prediction vector [15] [375] – Increase entropy of the prediction vector [16] [375] – Use regularization [17] [215, 375] • Differential privacy mechanisms [305]: – Differential privacy [18] [122, 435]. ◇ Adversarial regularization [19] [305] ◇ MemGuard [20] [204]

Table 2.2: Criteria and manifestations for *security*.

[9] It ensures data soundness by identifying abnormal input samples and by removing them [429].

[10] It ensures that algorithms are trained on statistically robust datasets, with little sensitivity to outliers [429].

[11] Adversarial samples are introduced to the training set [429].

[12] Input gradients are modified to enhance model robustness [429].

[13] The dimensionality of the network is reduced [429].

[14] Applicable when the number of classes is very large. Even if the model only outputs the most likely k classes, it will still be useful [375].

[15] It consists in rounding the classification probabilities down [375].

[16] Modification of the softmax layer (in neural networks) to increase its normalizing temperature [375].

[17] Technique to avoid overfitting in ML that penalizes large parameters by adding a regularization factor λ to the loss function [375].

[18] It prevents any adversary from distinguishing the predictions of a model when its training dataset is used compared to when other dataset is used [435]

[19] Membership privacy is modeled as a min-max optimization problem, where a model is trained to achieve minimum loss of accuracy and maximum robustness against the strongest inference attack [305].

[20] Noise is added to the confidence vector of the attacker so as to mislead the attacker's classifier [204]

Performance. The value of *performance* (see table 2.3) is defined by the (1) correctness of predictions [132, 138], along with the (2-5) resources necessary to reach such predictions [4, 61, 236]. The conditions under which systems are evaluated will have a direct impact on the “appropriateness score” that these systems will obtain in the form of a quantifiable indicator [117]. In other words, if the level of performance is solely measured in terms of accuracy, regardless of the needed data, prerequisites will be inherently favoring big “data-hungry” [241] models. As far as the measurement of performance is concerned, this is mainly done through quantifiable indicators, either referring to the preciseness of the results [294, 423] or to the estimated consumption of environmental resources [26, 104, 150, 151, 277].

Value	Criteria	Manifestations
Performance	1. **Correctness of predictions** [61, 132, 138] 2. **Memory efficiency** [4, 61] 3. **Training efficiency** [61] 4. **Energy efficiency** [4, 61] 5. **Data efficiency** [61]	• Accuracy (for classification, sum of true positive and true negative rates) [294, 423] • False Positive and False Negative rates [294, 423] • False Discovery and Omission Rate [294] • Mean and median error [423] • R2 score [60] • Precision and recall rates [423] • Area under ROC curve (AUC) [60] • Estimation of energy consumption through [151]: – performance counters – simulation – instruction- or architecture-level estimations – real-time estimation • Estimation of GPU memory consumption [150, 277] • Wall-clock training time [26, 104]

Table 2.3: Criteria and manifestations for *performance*.

Openness.
Transparency and explainability advocate for making system components and specifications accessible.

Transparency. Documents providing high-level principles for AI define *transparency* (see table 2.4) as the property that enables traceability and monitoring of AI systems [138, 298]. *Transparency* relates to the right to information [138] and requires that data or algorithms present some level of accessibility [391]. That is to say, data and models should present some level of (1) interpretability [61, 391], so as to (2) enable human oversight [138, 298]. Those data and models should also be (3) accessible [4, 138, 391], as a step towards achieving (4) traceability [298] and (5) reproducibility [61]. Manifestations of such criteria emerge mostly in the form of documentation detailing technical aspects of the AI system (considered signifiers in our framework) [4, 50, 83, 152, 153, 294, 298, 391]. Process-oriented practices mostly focus on giving open access to data and algorithms [4, 61, 138, 391], regularly reporting key information about the system [138] and notifying users whenever they are being subject to or interacting with an AI system [138].

Value	Criteria	Manifestations
Transparency	1. **Interpretability of data and models** [61, 391] 2. **Enabling human oversight of operations** [138, 298] 3. **Accessibility of data and algorithm** [4, 138, 391] 4. **Traceability** [298] 5. **Reproducibility** [61]	• Description of data generation process [4, 50, 83, 152, 153, 298] • Disclosure of origin and properties of models and data [4, 294, 391] • Open access to data and algorithm [4, 61, 138, 391] • Notification of usage/interaction [138] • Regular reporting [138]

Table 2.4: Criteria and manifestations for *transparency*.

Explainability. Explainable Artificial Intelligence (XAI) is formed by a set of techniques that allow a wide range of stakeholders to understand why or how a decision was reached by an AI system [141, 391]. *Explainability* (see table 2.5) is, thus, conceived as an interface that translates reasoning mechanisms of the system into formats that are (1) comprehensible [42, 61, 132, 138, 141, 142, 315, 391]. In addition, strategies for making black-box algorithms more interpretable facilitate their (2) monitoring [298] and, therefore, make them (3) suitable for evaluation [138, 298]. XAI techniques (process-oriented practices) are very diverse in nature. As claimed by Vera Liao et al. [255] and Barredo-Arrieta et al. [42], *explainability* methodologies are usually classified by the scope of the explanation, complexity of the model, model specificity and the stage of the AI pipeline where such a strategy is to be used. For our framework, we will consider that explainable models can be either (a) interpretable by design or they can be (b) explained by additional *post-hoc* explanations [42].

Value	Criteria	Manifestations
Explainability	1. **Ability to understand AI systems and the decision reached** [61, 132, 141, 142, 315, 391] 2. **Traceability** [298] 3. **Enable evaluation** [138, 298]	• Interpretability by design [42] • Post-hoc explanations [42]

Table 2.5: Criteria and manifestations for *explainability*.

2.3.4. Universalism vs Individual Empowerment

The second dimension captures the conflict between *universalism* and *individual empowerment*. Values included within the *individual empowerment* category emphasize the defense of the decision subjects' interests. These principles advocate for giving decision subjects the means to oppose to the conclusion reached and uphold the need for putting humans in the loop. Values within the *universalism* category emphasize the need to equalize system behavior to *all* and to ensure that such a system adheres to the interests of society as a whole, beyond the interests of a few individuals.

Universalism.
Respect for public interest, fairness and non-discrimination uphold the need to ensure equitable and socially acceptable system behavior for *all*.

Respect for public interest. The value of *respect for public interest* (see table 2.6) deals with the (1) appropriateness of developing AI systems for a certain purpose within a specific context. As Keyes et al. [218] claimed, making AI systems fairer, more transparent and more accountable is insufficient if we ignore the purpose of developing and implementing these systems in a certain context in the very first place [227, 397]. AI systems should, therefore, (2) be beneficial to society and humanity as a whole [138, 141, 142, 298], respect law [61] and be aligned with human norms [138]. This involves giving a clear justification of the purpose and benefits of building such a system [1, 83, 218, 298], so that the deployment of the system in question upholds public-spirited goals [138]. Universalism aims at protecting the welfare of *all*, both people and nature [236]. AI systems' (3) negative impacts on environment should, therefore, be considered and valued [4, 49]. To this end, process-oriented practices include the creation of diverse and inclusive forums for discussion [138, 292], whereas signifiers include the qualitative measurement of social and environmental impact [49, 298, 334].

Value	Criteria	Manifestations
Respect for public interest	1. **Desirability of technology** [1, 83, 218] 2. **Benefit to society** [138, 141, 142, 298] 3. **Environmental impact** [4, 49]	• Diverse and inclusive forum for discussion [138, 292] • Measure of social and environmental impact [49, 298, 334]

Table 2.6: Criteria and manifestations for *respect for public interest*.

Fairness. The value of *fairness* represents a complex concept that accepts multiple definitions [30, 217], some of which cannot be satisfied simultaneously [171, 181, 217]. Overall, we will understand *fairness* (see table 2.7) in terms of parity in output [116] and equal treatment [4] among individuals. When addressing more specific definitions of *fairness* (1-8), we will adopt the approach followed by Verma et al. [414], which was also echoed by Mehrabi et al. [284] (for a detailed enumeration and explanation of each of the definitions, the reader is encouraged to check table 2.7). ML techniques within AI systems generally conceive fairness in terms of statistical metrics [161] and observe whether specific quantifiable indicators are above or below the thresholds set for a certain application. Even if error rates were equal across groups for a certain application, if those rates are too high, the system could still be considered unfair [171]. This means that for our value-based framework we outline the quantifiable indicators that are normally used for manifesting fairness-related criteria, but we do not determine the threshold for these indicators to be considered good enough for a specific application. Similarly, the quantifiable indicators relate to the output of the system, rather than the outcome that these outputs lead to.

Value	Criteria	Manifestations
Fairness	1. **Individual fairness** [21] [42, 120, 237, 284] 2. **Demographic parity** [22] [42, 120, 171, 181, 216, 237, 284, 381, 414] 3. **Conditional Statistical parity** [23] [284, 414] 4. **Equality of opportunity** [24] [53, 170, 284] 5. **Equalized odds** [25] [284] 6. **Treatment equality** [26] [52, 284] 7. **Test fairness** [27] [86, 284, 414] 8. **Procedural fairness** [28] [165, 237, 284]	• Accuracy across groups (for classification, sum of true positive and true negative rates) [86, 171, 223, 298] • False positive and negative rates across groups [86, 223, 284, 349, 419] • False discovery and omission rates across groups [294, 349] • Pinned AUC [115, 294] • Debiasing algorithms [47] • Election of protected classes based on user considerations [165]

Table 2.7: Criteria and manifestations for *fairness*.

[21] Similar individuals should be treated in a similar way. Diverging definitions state that: two individuals that are similar with respect to a common metric should receive the same outcome (*fairness through awareness*); or any protected attribute should not be used when making a decision (*fairness through unawareness*); or the outcome obtained by an individual should be the same if this individual belonged to a counterfactual world or group (*counterfactual fairness*) [284].

[22] The probability of getting a positive outcome should be the same whether the individual belongs to a protected group or not [284].

[23] Given a set of factors L, individuals belonging to the protected or unprotected group should have the same probability of getting a positive outcome [284].

[24] The probability for a person from class A (positive class) of getting a positive outcome, which should be the same regardless of the group (protected group or not) that the individual belongs to [284].

[25] The probability for a person from class A (positive class) of getting a positive outcome and the probability for a person from class B (negative class) of getting a negative outcome should be the same [284].

[26] The ratio of false positives and negatives has to be the same for both groups [284].

[27] For any probability score S, the probability of correctly belonging to the positive class should be the same for both the protected and unprotected group [284].

[28] It deals with the fairness of the decision-making process that leads to the outcome in question [165].

Non-discrimination. The value of *non-discrimination* (see table 2.8), as defined in our framework, deals with AI systems not being socially biased [61] and ensuring that equal accessibility is provided to all individuals [298]. This means that (1) quality and integrity of data should be evaluated and ensured [138, 153, 181, 298, 325] in order to prevent "socially constructed biases, inaccuracies, errors, and mistakes" [298] from being present in the data. Processes that safeguard inclusive data generation [4, 83, 153, 298] and analysis procedures for identifying potential biases in data and for assessing its quality [138, 153, 181, 298, 325] are strategies that avoid social stereotypes being codified, maintained and amplified [181]. Furthermore, non-discriminatory systems should (2) ensure diversity and inclusiveness in the design process [132, 138, 298]. From a process-oriented perspective, participants involved in the development process should, thus, present diverse profiles [4, 138, 250, 441]. Finally, giving (3) equal access to the technology [4, 61, 138, 298] avoids the growth of inequalities as a consequence of deploying AI systems [138].

Value	Criteria	Manifestations
Non-discrimination	1. **Quality and integrity of data** [138, 153, 181, 298, 325] 2. **Inclusiveness in design** [132, 138, 298] 3. **Accessibility** [4, 61, 138, 298]	• Inclusive data generation process [4, 83, 153, 298] • Analysis of data for potential biases, data quality assessment [4, 138, 152, 181, 284] • Diversity of participant in development process [4, 138, 250, 441] • Access to code and technology to all [4, 61, 138, 298]

Table 2.8: Criteria and manifestations for *non-discrimination*.

Individual empowerment.

Contestability, human control and human agency address the politics behind AI systems [24, 426] and deal with the issues caused by power imbalances [61, 83, 209, 269, 406].

Contestability. The value of *contestability* (see table 2.9) is defined as the value that ensures that users have the necessary information to (1) enable argumentation against conclusions reached by AI systems [10, 31, 132, 138, 209, 248, 269, 391]. This involves (2) empowering citizens [31, 132, 209] to investigate and influence AI [209], as part of a broader regulatory approach [269]. As a matter of fact, *contestability* has been identified as a "critical aspect of future public decision-making systems" [10]. This implies that, from a documentation perspective (signifiers), users should be made aware of who determines what constitutes a contestable decision, who is accountable for it and who can contest a decision. This last point is particularly necessary to determine whether (legal) representatives of decision subjects can act on their behalf. The review mechanism in place and the workflow of contestations [269] are policy-related details that users should also be informed about. From a process-oriented standpoint, mechanisms for users to ask questions and to record disagreements should also be put in place [185, 295].

Value	Criteria	Manifestations
Contestability	1. **Enable argumentation / negotiation against a decision** [10, 31, 132, 138, 209, 248, 269, 391] 2. **Citizen empowerment** [31, 132, 209]	• Information of who determines and what constitutes a contestable decision and who is accountable [269] • Determination of who can contest the decision (subject or representative) [269] • Indication of type of review in place [269] • Information regarding the contestability workflow [269] • Mechanisms for users to ask questions and record disagreements with system behavior [185, 295]

Table 2.9: Criteria and manifestations for *contestability*.

Human Control. The value of *human control* (see table 2.10) addresses the influence that data-driven technologies have over humans and that leads to a reduction of human agency, power and control [324]. AI systems should be controllable [61] and (1) subject to user and collective influence [61, 248]. They should also be (2) subject to human review [138]. Governance mechanisms that ensure human oversight of automated decisions are, thus, necessary to maintain control and influence over such systems [298]. It should be possible to (3) choose how and even whether (in the very first place) to delegate a decision to an automated system [138]. From a development perspective, levels of human discretion should be established [132, 289] and the ability to override decisions made by a system [132] ought to be set up by design. Once the system is deployed, it should be continuously monitored to enable adequate intervention when necessary [132, 138, 389].

Value	Criteria	Manifestations
Human Control	1. **User/collective influence** [61, 248] 2. **Human review of automated decision** [138] 3. **Choice of how and whether to delegate** [138]	• Continuous monitoring of system to intervene [132, 138, 389] • Establishment levels of human discretion during the use of the system [132, 289] • Ability to override the decision made by a system [132]

Table 2.10: Criteria and manifestations for *human control*.

Human Agency. The value of *human agency* (see table 2.11) deals with the risks of AI systems displacing human autonomy [132, 138]. As claimed by Cila et al. [89], AI systems may displace human agency in governance processes and may undermine human autonomy. AI systems advocating for human agency should, therefore, (1) respect human autonomy [132, 138, 298] and (2) citizens' power to decide [61, 132]. In addition, (3) decision subjects should be able to opt out of an automated decision [132, 138]. The manifestations of such criteria involve giving knowledge and tools to users to comprehend and interact with AI systems [132] (signifier) and, from a process-oriented perspective, providing strategies for users to self-assess the systems [132].

Value	Criteria	Manifestations
Human agency	1. **Respect for human autonomy** [132, 138, 298] 2. **Power to decide. Ability to make informed autonomous decision** [61, 132] 3. **Ability to opt out of an automated decision** [132, 138]	• Give knowledge and tools to comprehend and interact with AI system [132] • Opportunity to self-assess the system [132]

Table 2.11: Criteria and manifestations for *human agency*.

Selecting values, criteria and manifestations for our example use case. Returning to the hypothetical insurance modelling team from our motivating example (section 2.3.2), they decided to apply our value-based framework before launching their system. They quickly realised that they need to consider more values than those outlined in current auditing processes. For example transparency, non-discrimination, supporting human agency and the public good. They also discovered a range of methods for enacting those values: from data handling processes that ensure anonymity and meaningful consent around the model, to models of fairness appropriate to their case.

Although we cover prominent ethical principles in AI, and the design and assessment of the AI system might include all of them, here we focus on a subset of those values for illustrative purposes. We imagine that the researchers developing the algorithmic life insurance application system want to focus on *explainability* and *privacy* (fig 2.2). We assume that they are dealing with a blackbox algorithm that is not interpretable by design. The team needs to examine whether the AI system and the decision reached are understandable. Additionally, the deployed XAI methods should enable traceability and evaluation of the system. As far as the *explainability* manifestations are concerned, since they are dealing with a blackbox algorithm, they need to deploy adequate post-hoc explanations. When it comes to *privacy*, the data used for training and testing the algorithmic model should have been obtained through the explicit approval of the decision subjects. These subjects should have been informed about the nature and purpose of the data that is collected, the way this data is handled and stored. Decision subjects should also have agency and control over their data. Additionally, data protection mechanisms should have been implemented to make sure that there is no possibility of identifying sensitive (in this case medical) data about the subjects. These two values that the team needs to advocate for, represent some trade-offs: XAI methods uphold interpretability of AI systems and some of them even rely on comparing data instances at inference time with those used for training the system. This would directly violate the subjects' right to have their data protected and confidentiality ensured.

2.4. Towards a multi-stakeholder reflection of AI systems

Since there are value trade-offs, like the one outlined in our example use case, and certain value-specific criteria are mutually exclusive, we follow the claim made by Raji and Smart [334], and advocate for standpoint diversity. This implies involving a wide range of stakeholders in the negotiation process [373] to discuss and critically reflect on the degree to which each of the values should be promoted in detriment of the other one and how the prioritization process should take place. These stakeholders will possess different types of knowledge and will present different insight needs. In this section we map those stakeholders and match them with the most suitable communication means.

2.4.1. Methodology for identifying relevant stakeholders and communication means

To identify relevant stakeholders, we follow the stakeholder characterization of Suresh et al. [386]. They classified stakeholders in a two dimensional matrix, where one dimension captured the nature of the knowledge of the stakeholders (formal, instrumental or personal) and the second one identified the context in which that knowledge manifests (Machine Learning, data domain, and the general milieu). Formal knowledge entails a deep understanding of the theories of a certain domain. Instrumental knowledge refers to the capability of applying formal knowledge in one of the three contexts. Personal knowledge is acquired by the participation of the subject in a specific context. The two dimensional-matrix classification results in nine different stakeholder profiles. To facilitate the process of mapping the stakeholders to tailored communication means, we narrow those stakeholders down into four categories [29].

We then proceed to identify the means to communicate system-related information to different stakeholders. We searched such means using arXiv and Google search, so as to cover the state of the art in terms of research papers and open source toolkits. Each search referred to specific value criteria and manifestations, although many of the found means address more than one value. This review does not intend to be exhaustive. We expect novel research to address value manifestations that still present scarce resources in our framework. Hence our review is just a snapshot of some of the available communication means until January 2022, but we host the latest version on an online repository[30] and is open to anyone's contribution. We aim at creating a living document that will keep growing and that will address current research gaps as time goes by.

2.4.2. Mapping stakeholders

We characterize four main stakeholders in our framework (see table 2.12): (1) The development team: they have the formal, instrumental and personal knowledge in the domain of AI [386]. They want to ensure and improve product efficiency and research new

[29] This reduced classification is backed up by the framework employed by Barredo-Arrieta et al. [42] when identifying the explainability needs of various stakeholders.

[30] https://github.com/mireiayurrita/valuebasedframework

functionalities [42]. (2) Auditing team: they have the formal and instrumental knowledge of the general milieu, meaning that they are aware of the social theories behind AI, and are able to evaluate technical specifications of AI systems. They aim at verifying model compliance with legislation [42] (3) Data domain experts: they have the theoretical (formal) and instrumental knowledge of the application context (healthcare, economics etc.). They look forward to gaining scientific or domain-specific knowledge [42, 386], trust the model [42, 386] and act based on the model output [386]. And (4) Decision subjects: they have the personal knowledge of the data domain in which the AI is being applied and the general milieu. They aim at understanding their situation [42], verifying that the decision is fair [42], contesting the decision (if needed) [386] and understanding how their data is being used [386].

Stakeholder	**Mapping** [386]	**Nature of knowledge**	**Purpose of insight**
Development team	ML, Formal + Instrumental + Personal	• "Knowledge of the math behind the architecture" [386] • "Stakeholder involved in an ex-ante impact assessment of the automatic decision system"[179]	• Ensure/improve product efficiency and debug [42] • Research new functionalities [42]
Auditing team	Milieu, Formal + Instrumental	• "Familiarity with broader ML-enabled systems" [386] • "Experts who intervene wither upstream or downstream" [179]	• Verify model compliance with legislation [42]
Data domain experts	Data domain, Formal + Instrumental	• "Theories relevant to the data domain" [386] • "Professional involved in the operational phase of the automatic decision system" [179]	• Gain scientific or domain-specific knowledge [42, 386] • Trust the model [42, 386] • Act based on the output [386]
Decision subjects	Data domain + Milieu, Personal	• "Lived experience and cultural knowledge" [386] • "Layperson affected by the outcomes of the automatic decision system" [179]	• Understand their situation [42] • Verify fair decision [42] • Contest decision [386] • Understand how one's data is being used [386]

Table 2.12: Description of potential stakeholders that can be brought together as part of our value-based framework. These stakeholders have been mapped following the two dimensional criteria (type of knowledge —formal, instrumental or personal— and contexts in which this knowledge manifests —ML, data domain, milieu—) outlined by Suresh et al. [386]. The nature of their knowledge and the purpose of gaining insight for each of them have also been defined.

Mapping stakeholders in our example use case Going back to our example, once *explainability* and *privacy* have been broken down into specific criteria manifestations, the team needs to map the stakeholders who will take part in the assessment process (fig

2.2). Based on the mapping presented in table 2.12, the development team represents the stakeholders who have the knowledge of the math behind the system. An external auditing team will join the discussion to make sure that the model is aligned with current legislation. Since the algorithmic life insurance application system deals with medical data, the data domain experts will be represented by a medical team and a life insurance expert. Decision subjects will be laypeople who seek to understand and verify their situation with regards to data usage and the decision reached by the system.

2.4.3. Mapping tailored communication means

We then examine each of the reviewed means and identify their typology, the value manifestations that they cover and the stakeholders that can make use of it, as illustrated in table 2.13 for privacy dashboards. The objective of mapping value manifestations, stakeholder profiles and communication means is that of enabling a fruitful and informed discussion among stakeholders. We classify these means in three categories: (1) *Descriptive documents* (red), (2) *Design strategies* (cyan), and (3) *Ready-to-use tools* (brown). Appendix A.1. summarizes the rest of the communication means and maps them to value manifestations and stakeholders for whom such methods are suitable.

The stakeholders assigned to a specific communication means are based on the audience addressed by the original authors of such methodologies. In some cases, the characterization of the intended audience was not as granular as our stakeholder mapping and the authors merely differed experts in AI from non-experts. Based on the nature of knowledge that we assigned to each of the mentioned stakeholders in section 2.4.1, we considered that the development and auditing teams are able to understand technically formulated system details (experts) whereas data domain experts and decision subjects would require more accessible communication means (non-experts). Similarly, some of the communication means identified for *explainability* are suitable for any stakeholder, but the original authors formulated the post-hoc explanations with varying degrees of complexity, which should be taken into account when trying to deploy such strategies. If the target audience are data subjects, we echo van Berkel et al. [53] and Cheng et al. [84] and recommend to limit presentation complexity and to instruct participants throughout the session.

It should be noted that this mapping process represents a first step to making a wide range of stakeholders with different backgrounds understand each other. We are aware that communicating system-related information in a tailored way does not directly lead to the resolution of value trade-offs, and that design strategies are necessary for facilitating such conversations [176]. In any case, the exercise of resolving value tensions should be a communicative process, rather than a simple explanation [313]. However, the means used for communicating specifications of the system will play a key role in the dynamics that will take place in those sessions.

Assigning communication means to each stakeholder in our example use case. The life insurance researchers are now looking into appropriate methods for communicating values to different stakeholders (fig 2.2), so that they can develop a comprehensive plan that ensures both compliance and communication of values.

Based on the mapping presented in appendix A.1., for the value of *explainability* and

	Means	Value	Manifestations	Stakeholder				Application	Visual elements
				DT	AT	DE	DS		
[B]	Privacy dashboards	Privacy	• Description of what and why data is collected • Description of how data is handled				✓	Agnostic	• Timelines • Bar charts • Maps • Network graphs
		Human Agency	• Self-assessment of the system						
		Transparency	• Disclosure of properties of data						

Table 2.13: Illustration of how we mapped communication means with values, manifestations and stakeholders (DT = Development Team; AT = Auditing Team; DE = Data Domain Experts; DS = Decision Subjects). Privacy dashboards are tools (green) that allow users to interactively assess the collection and usage of their data. The rest of the reviewed communication means are characterized in appendix A.1..

its manifestations in the form of post-hoc explanations, the team can use various design strategies and tools as part of their assessment process. To facilitate the navigation of the available communication means, they first examine appendix A.1. to locate the type of means (tool, strategy, or documentation), values and stakeholders they are interested in. Once they select the codes associated to each communication means, they check the selected communication means to see whether the value manifestations in question are addressed and to explore the details related to those means. If the team working on the life insurance case prefers a ready-to-use tool over the description of design strategies for assessing *explainability*, they can use InterpretML [310] and especially the DiCE [299] functionality, (code [AC]) with the development and auditing teams to evaluate counterfactual examples. These counterfactual examples tell how input features should change in order for the output of the system to be different. That is to say, how the individual applying for life insurance should be different, physically, or when it comes to insurance or medical history, for them to accept the application (if the original output was a refusal). However, this tool might not be suitable for non-experts who are not familiar with AI-related concepts. In the life insurance use case, the medical and insurance team and the decision subjects should receive a description of how the output changes if a feature is perturbed, absent or present adapted to their insight needs. This can be done by describing the answers to the questions “Why, Why not and How to be that” for a certain output [255] (code [P]). As for *privacy* manifestations, the development and auditing teams can examine data collection and storage specifications through the Datasheet [152] associated to the dataset in question (code [K]). Special attention should be paid to the “Collection” and “Preprocessing/cleaning/labelling” sections. For decision subjects, iconsets [131, 189, 283, 343] (code [A]) and privacy dashboards [123, 135, 137, 180, 444] (code [B]) are means for them to explore how their data is being used. It should be noted that the cell that intersects between data domain experts and *privacy* is blank. Based on the characterization of stakeholders that we provided, data privacy-related matters are not directly linked to the purpose that data domain experts show when willing to explore AI systems. This is translated into scarcity of methodologies related to *privacy* manifestations that directly address data domain experts.

2.5. Discussion and future work

We discuss important aspects of our framework below.

Design choices for creating a value-based framework. We aim at examining values that characterize AI systems rather than the organizations responsible for these systems. Hence, we did not integrate accountability or responsibility as a value *per se* in our framework. We are aware that algorithms cannot be held responsible for the potential harm that they might cause [71, 181], and that in order to effectively deploy such systems, there is an urgent call for accountability [10, 434]. Likewise, we are aware of the need for rigorous frameworks that support accountability [193] and we consider that the act of conceiving a design and assessment framework itself answers to the need to evaluate and audit AI systems [138]. Nevertheless, we did not explicitly highlight the profiles of the people accountable for the system. We decided to follow Zhu et al. [443] and considered accountability as a governance issue. We do, however, believe that entities up the chain of command should be held accountable for the potential harm caused by AI systems [4, 138, 181, 269, 358]. It should also be noted that values and criteria presented in this chapter might not be unique [373]. We acknowledge current discussions in VSD about the shortcomings of pre-selecting values [107] and, hence, do not claim universality. Extension and modification of values is possible in our framework, but are subject to respecting continuity and opposition between values. Similarly, criteria and manifestations can be extended and subsets could be included to create situationally-specific versions of the framework. Since the aim of our framework is to encourage critical reflection [145, 403] and we identified some value manifestations that require additional communication means, we particularly encourage those context-specific adaptations to happen. Under no circumstances should the scarcity of communication means for certain values identified in our framework represent an excuse to justify inaction or to ignore such values.

Context dependence and consistency. As echoed by Liscio et al. [261], in order to translate values into system requirements [329, 330], to reason about conflicting values [6, 302] and to communicate them to different stakeholders [146], it is necessary to situate these values within a context. The prioritization of values depends on the application context of such systems [4]. In this chapter, we showed an example of how the framework could be applied to a particular use case. However, considering the differences between value alignments and tensions that may arise due to context dependence, the validity and consistency of our framework is still to be tested. Future work needs to validate our framework across scenarios [53, 156] through user studies or synthetic experiments [379].

Need for standardization. To systematically review and revisit value priorities and tensions among different stakeholders, our framework should be part of a broader evaluation workflow [245], such as the one suggested by Raji and Smart [334]. Besides, practices from software engineering such as the Values Dashboard [311] could be adopted [374, 392]. This dashboard promotes awareness of values and aims at triggering discussions among stakeholders. It claims to be beneficial in each phase of the software development process, from inception to release, and establishes strategies, such as Timelines or Issues, that are already common practice on software development platforms like Github.

Implications of our work. Our multi-stakeholder value-based framework facilitates the unveiling of assumptions that encode political and social values made by developers [334]. By bringing together a wide range of stakeholders to evaluate and discuss value manifestations, one can anticipate and remedy harmful algorithmic behaviors before deploying a system. Besides, we provide researchers and industry practitioners with a good coverage of values to evaluate their systems and the association of such values to actionable value manifestations. This contributes greatly to the adoption of ethical approaches by practically-minded people [298]. For researchers, we provide them with an easy-to-navigate mapping of value manifestations, stakeholders and communication means. Our framework also visually illustrates research gaps that need to be addressed. Blank spaces in appendix A.1. or values with a scarce number of associated communication means directly refer to valuable research opportunities. For instance, for the value of *fairness*, a great deal of effort has been devoted to designing ready-to-use tools for stakeholders with a deep understanding of AI (developers and auditing teams). However, means for addressing *fairness* manifestations and communicating them to decision subjects have not received the same attention. For industry practitioners, we gathered ready-to-use open source toolkits that can be directly applied to their own use cases. Moreover, since we host this mapping on an online repository [31] open to future contributions, we hope that the number of tools addressing each of the identified value manifestations will grow and that the benefits of designing such a framework will be even more tangible in the future.

2.6. Chapter Takeaways

In this chapter, we developed a framework summarizing prominent ethical principles for trustworthy Artificial Intelligence design and assessment. We arranged eleven prominent ethical principles in a circular composition, so that common motivations and tradeoffs can be easily identified. We identified contestability as one of the ethical principles composing the “normative core” of a principle-based approach to trustworthy AI [138].

We then broke down each of these principles into a set of criteria and their correspondent manifestations in the form of quantifiable indicators, process-oriented practices, and signifiers. In addition, we examined available tools for communicating principle-specific manifestations to different stakeholders based on the nature of their knowledge and their insight needs. Through this mapping, we identified a scarcity of available tools for enabling a multi-stakeholder deliberation about contestability. Our findings in this chapter motivate further research into ways of operationalizing and deliberating about contestability, so that conflicting priorities can be reviewed, negotiated and consolidated.

[31] Check footnote 11

3

Decision Subjects' Needs for Contestability

In this chapter, we generate empirical insights into decision subjects' information and procedural needs for meaningful contestability (**RQ2**). To this end, we chose an illegal holiday rental detection scenario as our case; a high-risk decision-making process in the public sector. We conducted 21 semi-structured interviews with citizens with experience renting their homes out and different levels of AI literacy. We found that decision subjects request interventions that facilitate (1) cooperation in sense-making, (2) support in contestation acts, and (3) appropriate responsibility attribution. Our results highlight the cooperative work behind contestability, and motivate future efforts to structure individual and collective action, to personalize explanations for contestability, and to open up sites of contestation in AI pipelines.

This chapter is based on the following manuscript: **Mireia Yurrita**, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, Alessandro Bozzon. "Personalize, Prioritize, Collectivize: Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability". In *Proceedings of the ACM on Human-Computer Interaction CSCW* (CSCW '25).

3.1. Introduction

Several artificial intelligence (AI)[1] systems employed for decision-making in the public sector (e.g., AI for policy enforcement or for essential public services) [134] can negatively affect decision subjects' safety and fundamental rights, and are, therefore, considered *high-risk* by the European Union's Artificial Intelligence Act (EU AI Act) [133]. In order to safeguard decision subjects' rights to dignity and autonomy in high-risk algorithmic decision-making, an increasing number of scholars in the HCI community (e.g., [7, 406, 407, 437]) have claimed that AI systems should be *contestable* i.e., open and responsive to human intervention throughout their lifecycles [9]. Despite recent interest in making AI systems–and the decision-making processes where these are embedded–contestable, most prior work is theoretical, and has rarely accounted for the perspective of decision subjects when suggesting contestable AI design guidelines [9, 226, 355]. Failing to generate empirical insights into decision subjects' needs for contestability might, in turn, lead to designs that do not contribute to decision subjects' perceptions of control and voice [406, 437].

From a procedural perspective, the few empirical studies conducted to date have either (1) considered the standpoint of *human controllers* (i.e., domain experts who interact with the algorithmic system [9]) for identifying the challenges of implementing contestable AI systems in the public sector or (2) have focused on designing for contestability in contexts other than the public sector (e.g., content moderation [407]) [7]. The extent to which those findings are aligned with decision subjects' procedural needs for contestability in the public sector is unknown. From an information[2] perspective, recent work has explored the interplay between output explanations and *recourse* (i.e., operationalization of contestability that allows decision subjects to change the decision output by acting on input variables [404]). However, decision subjects might want to contest not only the decision output, but also more fundamental issues regarding the system (e.g., goal of the system, the idea of automation itself [406], or data sources [21]). It is still unclear which information enables decision subjects to engage in such contestation acts.

In this chapter, we aim to generate *empirical* insights into the procedural and informational means that *decision subjects* need to meaningfully contest *high-risk* public decision-making processes. Consequently, we seek to answer the following research question:

[1] Due to the "demand for data, technical complexity, and unpredictable interactions" [433] of AI systems, human-AI interactions are uniquely difficult to design for. This same nature of AI makes contestability uniquely difficult to design for [9].

[2] We will use the term (1) *information* to refer to a set of facts that describes a decision or a decision-making process, (2) *information item* to refer to a unit of relevant information [326], and (3) *explanation* to refer to tools or processes that an agent (explainer) uses to describe the decision (or the decision-making process) to another agent (explainee) [290]. An explanation involves a communicative effort for making the information that composes the explanation *understandable*. *Information needs* can, therefore, involve both *information items* —relevant content— or *explanations* —information (items) presented as part of an interaction.

What are decision subjects' information and procedural needs to meaningfully contest algorithmic decision-making processes?

To address this research question, we opted for a scenario in public decision-making; more specifically, a risk scoring system for the detection of illegal holiday rentals[3] (Section 3.3). We conducted 21 semi-structured interviews with participants who have personal experience renting out their homes as short-term rentals. We presented a scenario to our participants where they were detected by the algorithmic system, and asked questions on *what* they would like to contest in the decision-making process, *how* they would formulate their contestation, and the information they would need for it. Given the effect of AI literacy on users' information needs [219, 437], we ensured diversity in levels of AI literacy among participants. Our study was preregistered before data collection.[4]

Our results indicate that contestability in algorithmic decision-making is not limited to individual appeal processes, and requires a cooperative effort between *civil servants* (in roles that go from policy-making to AI development or street-level bureaucracy), *citizens*, and *third parties* (e.g., legal counsellors). As far as *information needs* are concerned, participants sought information that could help them make sense of algorithmic decisions and that would enable them take action to remedy the situation (Section 3.4.1). Participants expressed their willingness to engage in communication with human controllers and external parties to make sense of the provided information. When it comes to *procedural needs*, our participants expressed the need for support mechanisms (Section 3.4.2), i.e., they sought support both from the decision-making organization and from fellow decision subjects. Participants additionally highlighted the need for interventions that would ensure accountability in the decision-making process and social transparency (i.e., visibility of the complex socio-organizational context [126]) in public administration (Section 3.4.3).

In this chapter, we make two main contributions to:

1. We adopt an empirical approach to contestability and generate insights into decision subjects' information and procedural needs for contestability in the public sector.

2. We draw implications for practice and for research. These implications encourage public agencies and the research community to account for the cooperative work behind contestability.

Supplementary materials associated with this chapter include the pre-registration document, screening survey, interview protocol, and prompts used during the interviews. These are all openly available in our repository for the benefit of the community and in the spirit of Open Science (https://doi.org/10.4121/be171486-fe03-45fe-8d8b-22b4c81cd3a2).

[3] https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/ (last accessed 14.01.2024). Note: the entry of this algorithmic system in the algorithm register is from 2020. Due to delays in data collection as a consequence of the COVID-19 pandemic, the system has not been deployed. See Section 3.3.1 for information about the status of the system and the rationales behind choosing this case.

[4] https://osf.io/ejyt5 While more widespread in quantitative studies, by pre-registering our qualitative study we aim to (1) describe the original aims of the study, (2) register the assumptions that underlie the collection and analysis of the data, and (3) enable the scientific community to monitor the evolution of the study [173].

3.2. Related work

This section summarizes previous work on algorithmic decision-making in the public sector and contestable AI. In Section 3.2.1, we summarize prominent work on public AI. In Section 3.2.2, we include papers that have theoretically defined procedural means for contestability. In Section 3.2.3, we include literature concerning information needs for meaningful contestability.

3.2.1. Algorithmic Decision-Making in the Public Sector

In an environment like the public administration where decisions are often made based on incomplete, contradictory, and changing information [87, 260, 340, 376], the usage of AI has the potential to improve both the efficiency and quality of decision-making processes [447]. However, the development and use of AI for public decision-making has also been claimed to be uniquely challenging [356] mainly because street-level bureaucrats [13] need to be able to effectively apply human discretion while navigating bureaucratic processes in a resource-deficient context [356].

In addition to challenges in the development and effective use of algorithmic systems in the public sector, public AI systems face issues of perceived legitimacy [70]. Perceived legitimacy of public decision-making processes not only depends on the quality of the decision-making. According to the process-based model suggested by Tyler [399], the public's behaviour is "powerfully influenced by people's subjective judgments about the fairness of the procedure" through which decisions are made. Previous work has shown that communities impacted by algorithmic decisions in the public sector have concerns about the way in which data and algorithms are used [70]. While citizens are not opposed to delegating to fully autonomous systems, they do want to engage in a dialectical exchange with system controllers [12].

Contestability in algorithmic decision-making processes has, indeed, been defined as a dialectical exchange between decision-makers and decision subjects [355], a form of procedural justice that gives voice to decision subjects [9], and increases perceptions of legitimacy [317]. Perceived legitimacy of public decision-making processes, in turn, has been claimed to contribute to compliance, cooperation, and empowerment of citizens [399]. Given (1) the rapid adoption of AI systems in the public sector, (2) the potential (harmful) impacts of their widespread use, and (3) the relevance that *contestability* bears for procedural fairness and legitimacy perceptions in such high-stakes algorithmic decision-making processes, we decided to examine decision subjects' contestability needs in a public decision-making scenario.

3.2.2. Procedural Means for Contestability

Contestability refers to the quality that enables different actors (e.g., human controllers, decision subjects) to "*understand, construct, shape and challenge*" algorithmic decision-making processes [225]. Since algorithmic decision-making processes rely on interconnectivity [345] (i.e., the score that an individual gets is dependent on the scores of other individuals), designing ways in which decision subjects can meaningfully ensure a correct decision output and fair process is of paramount importance. Contestability has been conceptualized as *recourse* (i.e., the act of changing the output of an algorithmic system by altering input variables [404]), *appeal* (i.e., the act of opposing an algorithmic decision because it is considered to be faulty [412]) and as a design goal, *contestability by design* (i.e., AI systems that are open and responsive to human intervention throughout their lifecycles [9, 14, 355]). Both recourse and appeal are limited to acting on the decision output and are reactive in nature, whereas contestability by design allows measures to be taken *ex-ante* [9, 14]. Due to algorithmic systems' demand for data, technical complexity, and unpredictable interactions [433], contestability in algorithmic decision-making presents additional challenges compared to human-led decision-making [345]. Recent prominent work (e.g., [9, 14, 269, 355]) have set the grounds for conceptualizing contestability in algorithmic decision-making and have *theoretically* defined some procedural means that would enable algorithmic systems to be contestable by design.

Through a literature review, Alfrink et al. [9] synthesized five system features (e.g., built-in safeguards) and six development practices (e.g., agonistic development approaches) that contribute to contestable AI. Alfrink et al. [7] then used this framework to design a conceptual contestable AI system and identify the challenges of implementing contestable AI in the public sector. Similarly, Lyons et al. [269] analyzed responses to the Australian "AI Ethics Framework" which includes contestability as a key ethical principle and conceptualized how contestability could operate in relation to AI. Both frameworks were created based on theoretical claims without empirical grounding. There is, therefore, little insight into which of those elements decision subjects need to shape and challenge algorithmic decision-making. While acknowledging the importance of setting a normative framework that legally constrains the scope of contestability, in this chapter we argue that the lack of guidelines on decision subjects' procedural needs for contestability might result in contestation processes that either do not improve perceptions of legitimacy, in general [406] or that do not improve perceptions of *procedural voice* (i.e., ability to share one's views during a procedure [393]) and *influence* [393] in particular [437].

One of the very few *empirical* studies on decision subjects' needs for contestability was grounded in a context other than the public sector (i.e., content moderation [407]). The extent to which decision subjects' procedural needs for contestability in contexts such as content moderation can be extrapolated to contestability needs for contesting algorithmic decision-making in the public sector is not clear.

3.2.3. Information for Enacting Meaningful Contestability

For decision subjects to build arguments as part of their contestation process, they need knowledge, which, in turn, requires information [355]. This information needs to be meaningful for decision subjects to be able to engage in a rational and fruitful discussion [355], i.e., functional information that empowers decision subjects to exercise their *right to contest* algorithmic decisions as defined in Article 22(3) of the EU's *General Data Protection Regulation* (GDPR) [368]. Such information can be provided in the form of *explanations* [269] or *justifications* [179]. The goal of *justifications* is to demonstrate the appropriateness of the decision with respect to a norm (i.e., these are normative and extrinsic), whereas *explanations* aim at generating understanding about how a decision was made (i.e., these are intrinsic and factual). Information for meaningfully enacting contestability has been claimed to include the *why* behind the decision, as well as, *how* the decision-making process took place [9, 185, 355]. Despite the importance given to the topic, there is no empirical insights into the content and form of the justifications or explanations that decision subjects deem necessary. Determining what should or should not go into explanations is not trivial. Some decision subjects might want to "know everything" about *how* the system works, as it is the case for human-AI collaboration and for recommender systems [219, 370].

For contestability, previous work has mostly looked into generating decision output explanations for enabling decision subjects to engage in acts of *recourse* (e.g., [206, 336]). To this end, decision subjects need to *understand* [206, 416] and *act* [214] on an unfavorable decision through a set of *actionable* factors (i.e., factors that can be acted upon so as to change the decision output [213, 379]) or counterfactual explanations [206, 410]. Previous work indicates that when engaging in contestation processes, decision subjects might not only want to contest the decision output itself (scope of recourse) but also issues concerning the goals of the system or the idea of automation [406, 437]. Limiting information to output explanations might, therefore, hinder decision subjects' ability to question structural aspects (e.g., data sources [21]) of the decision-making process [162]. Current knowledge around *what* decision subjects would like to contest and *how* they would like to formulate their contestations might, therefore, be subject to blind spots resulting from limiting information to output explanations [269, 437].

3.2.4. Positioning Our Work

In this chapter, we aim to generate in-depth empirical insights into decision subjects' procedural and information needs for meaningful contestability that is not limited to algorithmic outputs. To this end, we conduct semi-structured interviews with potential decision subjects in a decision-making process in the public sector.

Our work builds on prior work and further informs it by:

1. **Adopting an *empirical* approach to identify needs for meaningful contestability**. Our results will provide insights into how decision subjects' needs align or differ from the claims made in theoretical frameworks for contestability summarized in Sections 3.2.2 and 3.2.3.

2. **Focusing on *decision subjects'* information and procedural needs for contestability**. Our results will provide a needs-based perspective that can further inform the organizational challenges for contestability identified by Alfrink et al. [7] in public administration.

3. **Focusing on a *public* decision-making context**. To this end, we choose a case in which risk scoring is used for fraud detection. Contestability needs that we identify might complement the ones identified by Vaccaro et al. [407] on content moderation processes.

3.3. Method

In this section, we introduce the case that we adopted for our study (Section 3.3.1) and summarize details about participant recruitment (Section 3.3.2), interview design (Section 3.3.3) and analysis procedure (Section 3.3.4). Note that the same interviews are also used for study 1 in chapter 5. In chapter 3 we analyze the interview data with a focus on identifying decision subjects' procedural and information needs for contestability. Instead, in chapter 5 the focus of the analysis is on identifying factors that impact decision subjects' perceptions towards different decision-maker configurations.

3.3.1. Case: Illegal Holiday Rental Detection

Algorithmic systems for law enforcement fall into the category of high-risk AI systems [133]. Within this category, we decided to select an algorithmic system suggested by the municipality of Amsterdam for accelerating the detection of illegal short-term rentals[5] as our case. The algorithmic system was designed to be used if a report on a particular address was received. After receiving the report, the algorithmic system (based on a random forest model) would compute the probability of a property being illegally rented for holiday purposes. It would do so by relying on data about the identity and housing rights of the decision subject, the building, and previous illegal housing cases. Based on the probability, civil servants would decide whether to further investigate the report. This system was suggested in November 2019 and expected to be pilot tested in 2020. However, due to the effect that the COVID-19 pandemic had on worldwide tourism, there were delays in data collection, which resulted in the system not being deployed to date (January 2024).[6]

Although the system has not been deployed, there are two main reasons why this represents a compelling case for identifying decision subjects' needs for contestability. First, this case deals with a timely and increasingly complex problem that impacts cities in several Western countries. Due to the issues that short-term rentals offered to tourists (e.g., Airbnb) have generated in the availability of long-term rentals for citizens [45], municipalities in several Western countries have started to regulate those rentals (e.g., Amsterdam, Barcelona) or even ban them (e.g., New York City) [308]. This last example is especially relevant. In September 2023, the municipality of New York City decided to ban short-term rentals that host more than two guests while the owner or tenants of the property are not present.[7] To enforce this policy, platforms like Airbnb are required to ensure their listings have pertinent licenses issued by the municipality certifying compliance with the regulation. This has led to the proliferation of a "black-market" where lessors use platforms such as Facebook or Craiglist to announce their short-term rentals and to avoid being policed by the platforms.[8] In response to this trend, many municipalities have put in place workflows where citizens can (anonymously) report an illegal holiday rental.[9] Algorithmic systems could, then, be seen as powerful tools to filter reports and help civil servants identify which reports they should investigate further. This is, precisely, the way in which the system suggested by the municipality of Amsterdam

[5] https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/(last accessed 14.01.2024)

[6] See the status of the project in the following official communication https://amsterdam.raadsinformatie.nl/document/12731876/2#search=%22Afhandeling%20toezegging%20pilot%20algoritme%20Alpha%20handhaving%20vakantieverhuur%22(last accessed 14.01.2024)

[7] https://www.nytimes.com/2023/09/05/nyregion/airbnb-regulations-nyc-housing.html (last accessed 14.01.2024)

[8] https://www.wired.com/story/airbnb-ban-new-york-illegal-listings/ (last accessed 14.01.2024)

[9] Barcelona: https://meet.barcelona.cat/habitatgesturistics/en; New York City: https://portal.311.nyc.gov/article/?kanumber=KA-02317; Berlin: https://ssl.stadtentwicklung.berlin.de/wohnen/zweckentfremdung_wohnraum/formular/adresswahl.shtml; Porto: https://www.asae.gov.pt/espaco-publico/formularios/queixas-e-denuncias.aspx (last accessed 14.01.2024)

was designed to operate. It is, therefore, a realistic representation of what municipalities in other Western countries could end up implementing. In an anticipatory exercise, the insights we get on decision subjects' contestability needs can be useful not only for the research community looking into contestability in algorithmic decision-making—*contestability by design*, which goes beyond post-hoc appeals, requires measures to be taken ex-ante [9, 14]—but also for municipalities thinking of implementing algorithmic systems to accelerate the detection of illegal holiday rentals.

Second, this case is part of the *algorithm register*[10] initiative launched by various European cities [139]. In an effort to ensure that algorithmic systems used for public services are "*responsible, transparent, and secure*", several cities (e.g., Amsterdam, Barcelona, Brussels) have put in place a register where information is provided about algorithmic systems used as decision support systems for public services. To this end, a short description about the system, information about mechanisms to ensure its responsible use, and technical information are openly shared. A form to provide feedback for continuous improvement is also included for each entry. The insights we get about decision subjects' *information* and *procedural needs* could, therefore, help improve a system that already advocates for transparency and contestability by design (e.g., by implementing mechanisms for quality assurance [9]).

3.3.2. Participant Recruitment and Selection

Given the timely and widespread applicability of the case (i.e., concerning major cities in several Western countries), we recruited participants who have experience renting out their homes as short-term rentals. Participants are located in municipalities from Western countries where workflows for detecting illegal holiday rentals have been put in place. Although it is unknown whether *all* these municipalities use algorithmic systems as part of those workflows (i.e., transparency around algorithmic systems used for public services is still not common practice [139]) if an algorithmic system like the one suggested by the municipality of Amsterdam was implemented, our participants could become decision subjects of the system by being correctly or incorrectly flagged. We recruited 21 participants in total (demographics in Table 3.1). We stopped collecting data when additional interviews failed to generate significantly new information. According to Clarke and Braun [90], when using qualitative interviews to capture experiences, understandings, and perceptions, the recommended dataset size is moderate (i.e., 10-20 participants), which aligns with the number of participants we recruited.

Since AI literacy has been shown to impact information needs [219, 437], we decided to ensure diversity in participants' AI literacy. We created a screening survey (cf. our repository) with questions about participants' literacy in and experience with AI. The screening survey comprised four items defined by Schoeffer et al. [365] (in a 5-point Likert scale). This way of operationalizing AI literacy has been used in prior studies and has been shown to be useful in capturing differences in informational fairness perceptions across individuals [12, 437]. We published the screening survey on online housing channels. We also put posters around our institution and reached out to personal contacts. We then selectively invited participants for our interview. To this end, we averaged the

[10] https://www.algorithmregister.org/ (last accessed 14.01.2024)

Table 3.1: Summary of our participants' demographics

Feature	**Category (Number of participants)**
AI literacy	High (7), Medium (7), Low (7)
Background	Computer Science (5), Engineering (4), Law (4), Business (3)[11], Design (3), Architecture (2), Physics (1), Social Work (1)
Country[12]	Netherlands (9), Spain (7), US (2), Portugal (1), Germany (1), Canada (1)
Immigration status[13]	Native (12), Non native (9)

four items that define AI literacy and divided participants in low, medium, and high AI literacy [12, 219, 220]. We reached out to participants while ensuring AI literacy diversity. As done in previous work [219], we refined the boundaries that define what constitutes low, medium and high AI literacy based on the interview answers given by our participants. This allowed us to account for potential discrepancies between self-assessed and functional AI literacy. A summary of our participants' AI literacy is provided in Table 3.2. We refer to our participants as P_k, where k is the identifier of a specific participant.

Table 3.2: Overview of our participants' AI literacy

AI Literacy	**Specification**	**Participants**
Low: self-assessment [1,3]	Had not heard much about AI	P14, P17
	Could not understand what AI entailed	P9, P13
	Unconfident about technicalities of AI	P3, P18, P19
Medium: self-assessment (3,4]	Technical background; familiar with basic statistics	P1, P2, P5, P11, P16
	Working on concepts adjacent to AI	P12, P20
High: self-assessment (4,5]	Working with or on AI on a managerial level	P6, P7, P10
	Working with or on AI from an engineering perspective	P8, P21
	Working with or on AI from a fairness perspective	P4, P15

3.3.3. Design of Interview Protocol and Materials

For our study, we opted to conduct qualitative interviews prompted by vignettes (i.e., written fictitious descriptions of events related to a topic of study [43, 351]). Choosing to run qualitative interviews allowed us to get rich and detailed insights into partici-

[11] Two of our participants have a joint background in Business and Law

[12] It refers to the country where the rented property is located. Our participants need to deal with that country's public administration for managing their property's rental.

[13] It refers to the mismatch between the home country of our participants and the country where the rental is located. Our participants are native or non native in the eyes of the public administration of the country where the property is located.

pants' needs for contestability [90]. As suggested by Clarke and Braun [90] when using qualitative interviews to capture participants' perceptions and needs, participants had a hypothetical personal stake in the selected case (i.e., they were renting properties as short-term rentals). The usage of vignettes has been claimed to be appropriate to capture perceptions, beliefs, and attitudes in social research, as well as to identify participants' reactions and needs in a particular situation [43, 192]. Scenario- or vignette-based techniques have previously been used in qualitative AI research; for instance, in public AI research for exploring the perspectives of decision subjects in the early stages of AI system usage in child welfare services [70], or in explainable AI research for advancing the conceptual development of *social transparency* [127]. Our interviews comprised four main sections and three prompts to encourage our participants to expand on their answers [90]. The interview protocol and prompts can be found in our repository and in Appendix B.1.

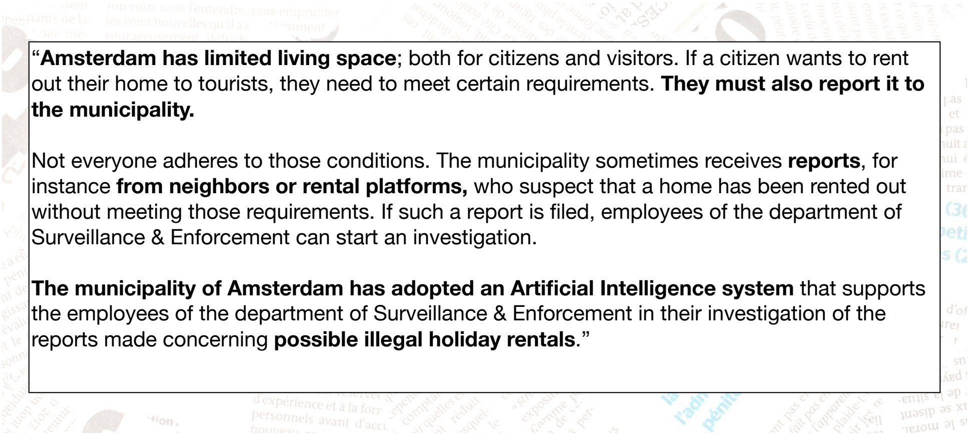
1. *Participants' background and experience.* First, we asked a series of questions to capture our participants' experiences with contestation processes. The objective was to identify their motivations for deciding whether or not to contest an unfair decision. We also asked them about their experience and motivation for renting out their homes.

2. *Perceptions around the use of AI.* We then introduced the first prompt (i.e., a fictional *piece of news* introducing the case; Figure 3.1), and asked our participants about the appropriateness and benefits of using algorithmic systems for detecting possible illegal holiday rentals. The fictional piece of news included real information about the system summarized from the introductory text in the algorithm register entry. The piece of news was tailored to the city where the rental was located to make the scenario more believable and for participants to feel they had a personal stake in the topic [90]. The objective of this section was to get a sense of how our participants perceived algorithmic systems (e.g., its perceived capabilities [207]) as a way to get context to their motivation for contesting (or not) the algorithmic decision-making process.

3. *Object of contestation (what to contest) and means for contesting (how to contest).* Next, we introduced the second prompt (i.e., a *letter*; Figure 3.2). The letter was divided into three main sections. These included (a) first warning and future penalty (i.e., giving notice [210]), (b) right to present arguments against the decision by calling the municipality (i.e., right to be heard [210]) and (c) right to know about the decision and the decision-making process (i.e., reason giving [210]). The amount of the penalty[14] and the timeframes for contesting[15] are informed by the contestation procedures available in the municipality of Amsterdam, within the Dutch public administration context. The letter was tailored to the city where the rental was located. For this interview, we deliberately designed the letter using accessible language (i.e.,

[14] https://www.amsterdam.nl/wonen-leefomgeving/wonen/boetes-overtredingen-vakantieverhuur-bed/ (accessed 14.01.2024)

[15] https://www.cjib.nl/direct-regelen/ik-ben-het-niet-eens-met-mijn-boete (accessed 14.01.2024)



“**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**.”

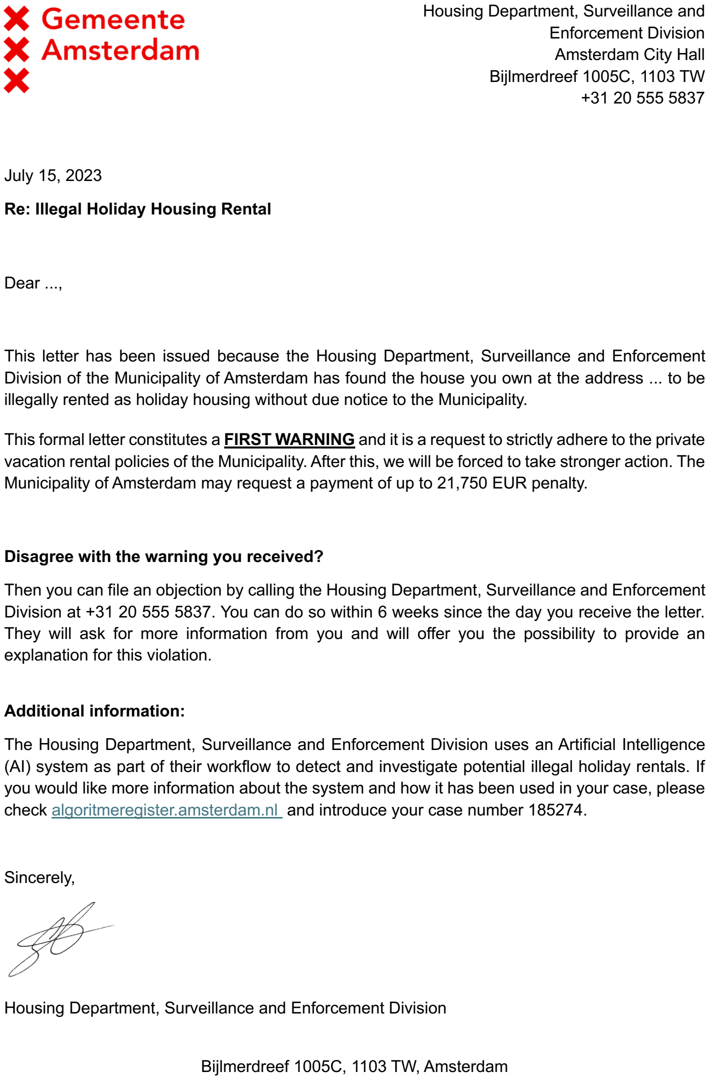
Figure 3.1: Example of the piece of news shown to participants to introduce our case. The material used with each participant included the name of the city where their short-term rental was located.

avoiding legal jargon), following the guidelines on accessibility of (digital) communications of public authorities.[16] We asked our participants how they would react to this letter and how appropriate they considered the contestation means (i.e., a phone call) suggested by the municipality (i.e., perceived voice and influence [393], expected treatment [56]). Through this section we aimed to capture *what* our participants would like to contest and *how* they would ideally like to proceed [270].

4. *Information needs.* Finally, we introduced the third prompt (i.e., the *information sheet*). The information available in the algorithm register was summarized in three categories [224] and organized through a color code (Figure 3.3): (a) green for information related to the scope of the system (i.e., reasons for system conception, role of the system and potential harms [275]), (b) orange for the decision rules of the process (i.e., information about training data [255, 275], system architecture [275]) and (c) blue for information related to the outputs (i.e., rationales behind instance-level decisions and model performance [255]). For the decision explanation, we simulated a SHAP explanation [267] (i.e., feature-based explanation [365])[17]. We indicated data features that contributed to the decision. We used positive (+) signs to indicate that a data feature contributed to high fraud risk [59, 116]. We avoided to include data features that are explicitly protected by law (e.g., gender [46]) in the decision explanation. Through this prompt, we asked our participants what they would like to know more about. The objective of this section was to identify if they would use this information for building their arguments as part of the contestation process [355, 416]. We decided to introduce information about the system after the letter to see if there

[16] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016L2102 (accessed 14.01.2024)

[17] The numbers in the presented SHAP value (first blue box in Figure 3.3) aim at representing the effect of each feature on the output risk value, rather than their effect on the final risk probability. That is why some of these features present an effect > 1. We left it up to the participants to ask for clarifications about the scale if they considered this information item to be important for contestability.



Gemeente
Amsterdam

Housing Department, Surveillance and
Enforcement Division
Amsterdam City Hall
Bijlmerdreef 1005C, 1103 TW
+31 20 555 5837

July 15, 2023

Re: Illegal Holiday Housing Rental

Dear ...,

This letter has been issued because the Housing Department, Surveillance and Enforcement Division of the Municipality of Amsterdam has found the house you own at the address ... to be illegally rented as holiday housing without due notice to the Municipality.

This formal letter constitutes a **FIRST WARNING** and it is a request to strictly adhere to the private vacation rental policies of the Municipality. After this, we will be forced to take stronger action. The Municipality of Amsterdam may request a payment of up to 21,750 EUR penalty.

Disagree with the warning you received?

Then you can file an objection by calling the Housing Department, Surveillance and Enforcement Division at +31 20 555 5837. You can do so within 6 weeks since the day you receive the letter. They will ask for more information from you and will offer you the possibility to provide an explanation for this violation.

Additional information:

The Housing Department, Surveillance and Enforcement Division uses an Artificial Intelligence (AI) system as part of their workflow to detect and investigate potential illegal holiday rentals. If you would like more information about the system and how it has been used in your case, please check algoritmeregister.amsterdam.nl and introduce your case number 185274.

Sincerely,

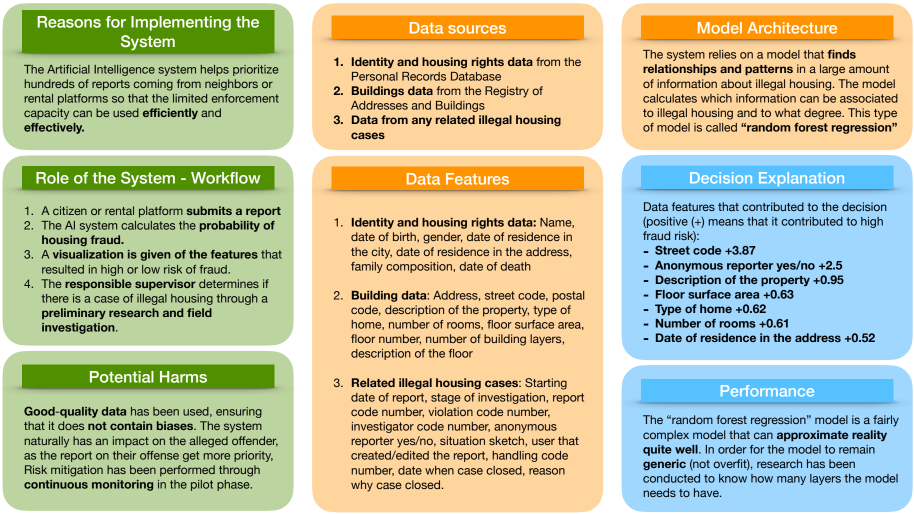
Housing Department, Surveillance and Enforcement Division

Bijlmerdreef 1005C, 1103 TW, Amsterdam

Figure 3.2: Example of the letter shown to participants. The letter used with each participant was tailored to include their name, address, the logo, name and contact details of the municipality where their short-term rental was located.

were any differences between the object of contestation before and after being given information about the system [162].

Ecological validity: We (the authors) designed the interview material so that (1) it



Reasons for Implementing the System

The Artificial Intelligence system helps prioritize hundreds of reports coming from neighbors or rental platforms so that the limited enforcement capacity can be used **efficiently** and **effectively.**

Role of the System - Workflow

1. A citizen or rental platform **submits a report**
2. The AI system calculates the **probability of housing fraud.**
3. A **visualization is given of the features** that resulted in high or low risk of fraud.
4. The **responsible supervisor** determines if there is a case of illegal housing through a **preliminary research and field investigation**.

Potential Harms

Good-quality data has been used, ensuring that it does **not contain biases**. The system naturally has an impact on the alleged offender, as the report on their offense get more priority, Risk mitigation has been performed through **continuous monitoring** in the pilot phase.

Data sources

1. **Identity and housing rights data** from the Personal Records Database
2. **Buildings data** from the Registry of Addresses and Buildings
3. **Data from any related illegal housing cases**

Data Features

1. **Identity and housing rights data:** Name, date of birth, gender, date of residence in the city, date of residence in the address, family composition, date of death

2. **Building data**: Address, street code, postal code, description of the property, type of home, number of rooms, floor surface area, floor number, number of building layers, description of the floor

3. **Related illegal housing cases**: Starting date of report, stage of investigation, report code number, violation code number, investigator code number, anonymous reporter yes/no, situation sketch, user that created/edited the report, handling code number, date when case closed, reason why case closed.

Model Architecture

The system relies on a model that **finds relationships and patterns** in a large amount of information about illegal housing. The model calculates which information can be associated to illegal housing and to what degree. This type of model is called **"random forest regression"**

Decision Explanation

Data features that contributed to the decision (positive (+) means that it contributed to high fraud risk):

- **Street code +3.87**
- **Anonymous reporter yes/no +2.5**
- **Description of the property +0.95**
- **Floor surface area +0.63**
- **Type of home +0.62**
- **Number of rooms +0.61**
- **Date of residence in the address +0.52**

Performance

The "random forest regression" model is a fairly complex model that can **approximate reality quite well**. In order for the model to remain **generic** (not overfit), research has been conducted to know how many layers the model needs to have.

Figure 3.3: Information Sheet provided to our participants. It includes the information relevant to the algorithmic decision-making process summarized from its entry in the algorithm register. It is color-coded. Green refers to information related to the scope of the system and it includes reasons for implementing the system, the role of the system, and potential harms. Orange refers to information about the decision rules of the system and it includes data sources, data features, and model architecture. Blue refers to information related to the outputs and it includes the decision explanations and performance information.

would illustrate a decision-making scenario where the illegal holiday rental detection system could be embedded, and (2) it would be sufficiently believable for our participants, i.e., it would not be considered science fiction [7, 40]. To improve the ecological validity of our work, we ensured participants were coming from cities where illegal holiday rental detection efforts are already in place and we tailored the materials (e.g., logos, address, recipient name) to the city where each participants' property was located. We additionally pilot tested the interview protocol and the prompts with 2 experts in human-computer interaction (different from the authors) from our institution. For each interview question, we evaluated whether it helped answer our research question, we looked for problematic assumptions, and we reflected on how meaningful participants would find it [90]. For each prompt, we checked the wording and layout. Based on the insights we got from the pilot test, we modified the layout of the *Information Sheet* to make it more engaging. We decided to change the decision explanation to textual form [53, 365], rather than a visual to avoid saliency bias and halo effect [118, 130].

3.3.4. Data Collection and Analysis

Data Collection We conducted the interviews between July and August 2023. All interviews were conducted online, using the Zoom video conferencing tool, and lasted 1 hour on average. Participants were offered 25 EUR (or equivalent) as compensation for their time. Our study was approved by a research ethics committee at our institution. All our participants signed an informed consent form. After each interview, we acquired the transcription of the recording through the videoconferencing platform if the interview was conducted in English. We then anonymized the transcription. If the interview was conducted in a language other than English, we had the recording transcribed in the original language through a third-party transcription software, anonymized and then locally translated the transcription using *DeepL*.[18] After obtaining the transcriptions in English, we reviewed and corrected them.

Data Analysis A critical realist [144, 276] and contextualist [276, 388] approach underpins our analysis. We acknowledge that although a reality exists and informs our findings, we, as researchers, play a role in constructing knowledge and these findings cannot, thus, be considered truly objective [424]. We analyzed our data using *reflexive thematic analysis* with a combination of inductive and deductive orientation to data [68, 91]. Reflexive thematic analysis is a flexible method that allows an in-depth engagement with the data. This approach is adequate for answering our research question where we aim to identify patterns *in data* and interpret them [67]. We conducted the data analysis on *Atlas.ti*.[19]

Analysis Procedure Data analysis was led by the first author. After transcribing –and translating when applicable– the recordings, the first and second authors cleaned the transcriptions. The first, second, and third authors read the transcriptions and got familiar with the material. The first author open-coded the transcripts and clustered the codes

[18] DeepL Translator: https://www.deepl.com/en/translator.

[19] Atlas.ti URL: https://atlasti.com.

in code groups. The second and third authors partially coded the data and reviewed the code groups. The first author then crafted the themes. All authors reviewed and mapped the themes. In total, three main themes and six sub-themes were developed. The first author finally refined the codes based on the final themes. Having different researchers analyze the data helped us reflect on different perspectives on the same data to develop richer insights into that data. Reflexivity helped researchers identify their own situatedness within the research and take responsibility for it [91].

Statement of Positionality. Reflexivity acknowledges that knowledge production is contingent on the researcher producing it [51]. As researchers living and working at a Western European university, we recognize that our perspectives shape the research and knowledge we generate. Our disciplinary backgrounds include engineering, cognitive science, computer science, HCI, and design. We have previously argued for making algorithmic decision-making processes contestable.

3.4. Results

The cooperative nature of contestability was a salient characteristic of contestation processes and was present throughout the interviews. We structure our results to highlight the cooperative work involved in contestability at three different points in time: (1) during the sense-making process that enables decision subjects to understand the provided information (post-hoc intervention[20]; Section 3.4.1), (2) during the contestation act (post-hoc intervention; Section 3.4.2), and (3) during the development and deployment of the AI system (ex-ante intervention; Section 3.4.3). We provide an overview of the themes and sub-themes in Table 3.3. We name themes as T*i.* and sub-themes as T*i.j.*, where i and j are the indexes of a particular theme and sub-theme. To improve readability, we avoid naming participants (P_k) for *each* statement that compose our themes and sub-themes. We, instead, give a sense of the prominence of each statement by using terms such as *a few, many, mostly, generally, unanimously.* A detailed mapping of the participants whose responses led to the statements in our results section is included in our repository. Additionally, we release our codebook, where we include the specific quotes that compose each statement. The codebook can also be found in our repository.

3.4.1. T1. Cooperation in Sense-Making

The first theme highlights the need for cooperation in the sense-making process that precedes the contestation act. **This cooperative effort involves decision subjects, legal and AI experts that decision subjects could contact, and street-level bureaucrats acting as controllers.** Controllers are street-level bureaucrats that are involved in the first instance of the decision-making and that interact with the decision subject to inform them about their situation before starting a contestation act. We observed an effect of AI literacy on decision subjects' information needs for sense-making (e.g., the type of information that participants with different levels of AI literacy perceived as actionable).

[20] We use the term *post-hoc* intervention to refer to an intervention that happens once the algorithmic decision is made. For interventions that happen before the algorithmic decision is made, we use the term *ex-ante* intervention.

Table 3.3: Overview of themes and sub-themes.

Information and Procedural Needs
T1. Cooperation in Sense-Making – post-hoc intervention
T1.1. Strategizing Information Requests
T1.2. Facilitating Dialogue with Controllers
T2. Social Support in Contestation Acts – post-hoc intervention
T2.1. Seeking For Organizational Support
T2.2. Seeking For Peer Support
T3. Distributed Responsibility – ex-ante intervention
T3.1. Ensuring Algorithmic Accountability
T3.2. Fostering Social Transparency

T1.1. Strategizing Information Requests

Participants in our study developed strategies for deciding which information to request. These strategies depended on participants' ability to make sense of the provided information –or their capacity to look for an independent expert advisor who could help them make sense of the information– and the risks and benefits of issuing an appeal using that information. Participants generally hypothesized two reasons for receiving the letter: (1) they had violated the regulation, or (2) they represented a false positive. In case (1), participants would accept the decision based on the justification and evidence they are shown. In case (2), there were two main situations that participants contemplated: (2a) they had rented their property out but they had a license for it, or (2b) they had not rented their property but the system indicated that they did.

In view of the above, participants, regardless of their AI literacy, unanimously prioritised knowing *why* they got detected by the algorithmic system. The reason behind this was that knowing *why* they got flagged was the first step towards knowing which of the situations they were in and resolving the issue. Participants pointed to the difference between a feature-based explanation, and a decision justification that clearly signals the reasons why a penalty is issued. The provided decision explanation (see Figure 3.3) did not satisfy their information needs because it did not provide a clear actionable path that could help decision subjects remedy the situation. P16, for instance, complained about the uselessness of receiving a feature-based explanation that points how the system had identified their property as an illegal holiday rental:

> *"Are you telling me that I have illegally rented my house or are you telling me there is a probability of me illegally renting my house? That probability could be based on a thousand things. Tell me the things I have actually missed. There has to be a concrete reasoning behind it, just give me that reasoning. Don't give me these numbers."* (P16)

Many participants additionally wanted to understand the decision basis (i.e., the policy behind the decision) to better discern whether their actions conform to the law or to double-check that the algorithmic decision basis was backed up by relevant policy. Some admitted that they might not have been aware of the regulation and would accept the first warning if this was duly motivated. The option of asking for legal advice to make sense of the decision and the policy was mentioned several times. The willingness to ask for legal advice depended on the required resources and the amount of the penalty.

Among the hypothesized scenarios, it was only in case (2b) that some participants started questioning the algorithmic system beyond the output itself. This was motivated by the difficulties in showing proof of innocence in this particular scenario as compared to the other scenarios. Our participants' AI literacy and experience dealing with fairness in AI affected their interest in knowing *how* the decision-making took place (including *how* the algorithmic system worked) and the perceived actionability of such information. This can be explained by the effect that participants' AI literacy had on the sense-making process that precede a contestation. Participants with *low AI literacy* were mostly uninterested in knowing how the algorithmic system worked because they were not certain about how they would use this information as part of their contestation. *Medium AI literacy* participants were interested in receiving more information about the data used by the system since this would allow them to ask questions related to privacy and bias in data. They were also generally curious to know more about the system due to their technical background but expressed doubts on how to use this information. For example, when asked about their willingness to know more about the AI system itself, P10 responded:

> *"Myself, because I'm quite a freak, I would [like to know more on a system level]. In general, I don't think people would care. They would be very focused on [fixing] their own problem."* (P10)

The option to contact experts in AI that would help them make sense and act upon information regarding the workings of the AI system was mentioned as an option by some participants with medium AI literacy. Among *high AI literacy* participants, the perceived actionability of AI-related information was further influenced by participants' experience in topics related to AI fairness. This was due to the effect that experience with AI fairness had on participants' ability to identify the subjectiveness of many of the design choices in the development of AI systems and their capacity to use this information as part of their contestation. Participants with high AI literacy who had not previously dealt with fairness-related topics were aligned with medium AI literacy participants and had doubts about how they could use information about the AI workings as part of the contestation. Instead, participants with high AI literacy and experience with AI fairness were willing to know and question aspects related to e.g., AI development. See table 3.4 for a detailed account of the information that each subgroup of participants deemed important and actionable. As shown in the table, our results indicate that the sense-making process that precedes contestation acts depends on decision subjects' AI literacy and AI fairness experience.

T1.2. Facilitating Dialogue with Controllers

Responses from our interviews indicate that the way in which decision subjects make sense of the provided information is also influenced by the means used for dialogue between decision subjects and human controllers. The communication between controllers and decision subjects turns information into meaningful explanations. This communication also helps clarify technical jargon, a key aspect of the sense-making pro-

AI literacy	**Information needs - perceived as actionable**	**Contestation object - *what* to contest**
Low AI literacy	• *Why* of decision output	• Correctness of decision • Legal validity of AI usage • Appropriateness of decision basis • Lack of legitimate proof
Medium AI literacy *High AI literacy* (no experience in AI fairness)	+ Data-related information	+ Bias + Discrimination + Privacy of data
High AI literacy (experience in AI fairness)	+ Decision explanation + Model-related information + Development of the system	+ Explanation weights + Prioritization of data features + Lack of model robustness + Faulty development process

Table 3.4: Overview of the relationship between AI literacy, information needs and contestation objects. + symbols indicate that the presented items are cumulative (e.g., medium AI literacy participants wanted to know the *why* behind the decision output as well as information related to data). This arrangement shows tendencies we observed in our data and does not necessarily represent one-to-one relations: not all participants from a given subgroup requested all of the information and a few participants requested more information than it is indicated for their subgroup. For a one-to-one relation for each item, see the mapping in the supplementary material.

cess according to our participants. One aspect that impacts the communicative effort and joint sense-making between controllers and decision subjects is the communication channel. Participants' preference for communication channels varied based on the stakes and complexity of the decision, who was responsible for the situation, personal experience, and language. However, there was consensus that communication channels should be designed to minimize the friction of engaging in a dialogue.

When it comes to the format in which information is presented, almost all participants wanted the effort to understand the information provided by the municipality to be minimal. This required the provided information to be relevant to their case, concise, simple, and clear. The reason behind this was the need to make the information digestible to different decision subjects, especially those with lower levels of AI literacy. A potential means for satisfying this requirement would be the progressive discovery of information based on relevance. As claimed by P8, this could be achieved through information hierarchies.

> *"Give me a diagram of my case. Then if I want to go in detail on anything in particular, put it all at the end. I would love for everything to be well explained at the end."* (P8)

Many deemed visual explanations (i.e., graphics) or explanatory videos of the decision-making process as appropriate mediums of communication due to their interactivity.

3.4.2. T2. Social Support in Contestation Acts

The second theme focuses on the contestation act itself. **Contestation acts require the joint effort of decision subjects, street-level bureaucrats acting as reviewers, third parties assigned to decision subjects, and fellow sufferers.** Reviewers are street-level bureaucrats involved in the contestation process [271]. Contestation acts were defined to participants as processes implemented *within* the organization rather than before a tribunal [354]. For this reason, cooperation with legal representatives was not mentioned as a central element of the contestation process (unlike the sense-making). We observed an effect of AI literacy on decision subjects' procedural needs for engaging in acts of contestation (e.g., the need for third parties to compensate for knowledge differentials).

T2.1. Seeking Organizational Support

Reviewers were seen as key actors in supporting decision subjects on a one-to-one basis and in facilitating the act of contestation. Our participants generally preferred a human reviewer over an algorithmic reviewer. The reasons for this were varied. AI was seen as unable to change the output. Humans, in contrast, were seen as more appropriate since they could provide answers beyond frequently asked questions, and they could deal with *grey* areas (i.e., ill-defined situations).

Our participants expressed a general wish for the human reviewer to be cooperative and empathetic during the discussions. Many highlighted the need for a proactive attitude where *"both parties need to be willing to find a solution and a conclusion to the problem"* (P12). The reviewer thus needs to be an active listener as opposed to a *"nameless bureaucrat who doesn't really deal with my issue."* (P7). The reason for this was the wish of decision subjects to feel understood and *"to be heard before being given a warning."* (P21). In contrast to reviewers in existing public decision-making processes, many participants claimed that reviewers for algorithmic decision-making should be experts in AI so that they can effectively accompany them throughout the contestation process.

While participants acknowledged the need for cooperation with human reviewers, a few defined the contestation process as a fight. One of the main reasons for this was the *power differentials* between the decision subject and the reviewer. Power differentials are accentuated with knowledge differentials (e.g. when decision subjects have low AI literacy or when non-native decision subjects do not know the functioning of public administration). Many participants requested a third party (e.g., a watchdog) to mediate the conflict. The third-party could ask questions on decision subjects' behalf and could have information about similar cases. P2, for example, motivated the need to have an independent party involved in the contestation process to deal with information differentials:

> *"I would want a third party. Someone who is equally informed but who did not build the system. Just to have an objective perspective."* (P2)

The third-party should have both legal and technical knowledge (i.e., experience in data science) and should help decision subjects to move forward. A few participants acknowledged that the level of support needed from the third party would depend on the decision subjects' AI literacy, the level of satisfaction with the dialogue they had with the controller during sense-making, and decision subjects' legal knowledge.

T2.2. Seeking Peer Support

Participants generally prioritized clarifying their own case at the individual level—for all scenarios (1), (2a), and (2b) in section 3.4.1. The possibility of contesting aspects of the algorithmic system—scenario (2b)—, however, was conceived to be more feasible if done collectively. When asked about the possibility of contesting the algorithmic system, P11 mentioned:

> *"If more of us get this letter, then maybe a consortium could be formed, and then through that consortium, we would discredit the AI system. If I was the only one of my social circle getting this letter, I wouldn't immediately go towards discrediting their AI. "* (P11)

Some suggested that using similar cases where the algorithmic system repeatedly made an error could be the basis of the collective contestation. This would be a means for others not to go through the same issues if the system incorrectly flags them. The collective was regarded as *"a place that is organized by citizens, by people that have gone through this"* (P18). Within those previously affected, high AI literacy individuals, or experts with some status (e.g., professors) could be the technical guides to help escalate the situation. According to P6 and P15, attracting the attention of the media and turning the issue into a political matter would be required.

> *"[If] there is a group that we all together try to say [that] this shouldn't work like that, and this becomes a thing, then it could form a very interesting, small nerdy rebellion against the AI system. I think that this is a collective issue, which needs space and people, and attention. (...) It can be like an Anonymous kind of thing, but for AI and for governmental AI systems."* (P15)

Some participants highlighted that a collective could help citizens affected by the system to remedy their situation. A collective would provide decision subjects insights into similar cases. They claimed that this could also enable spotting of error patterns across false positives. This was deemed especially important for people with *low AI literacy* and with no immediate social support structures for providing emotional and procedural help.

3.4.3. T3. Distributed Responsibility

The third theme highlights the **need for street-level bureaucrats acting as controllers, policy makers, and other members of the public administration to cooperate and to ensure appropriate responsibility attribution.** We did not observe an effect of AI literacy on claims about responsibility attribution.

T3.1. Ensuring Algorithmic Accountability

In general terms, participants appreciated and wanted to exercise their right to contest the algorithmic decision but dealing with the consequences of errors made by the algorithmic system was perceived to be unfair. Many mentioned the burden of showing proof of innocence and the effort needed to make sense of the information that would enable them to do so. Overall, there was consensus on the fact that correcting AI's mistakes is not the decision subject's responsibility. If such a burden is put on the decision subject and this represents a false positive, a few participants requested *compensations* for the time wasted and the effort devoted to contesting.

There were several views on whose responsibility it was to contest the system. P15, for example, mentioned that, *"what I would like is the AI to be contested by the employees before they send you the letter."* This would require human controllers to be able to identify such false positives, for which P21 suggested an approach. The suggested workflow would entail: (1) the municipality contacts the individual that has been flagged by the system before any warning is issued, (2) the municipality provides the reasons why they contact the individual, (3) there is a discussion around the reasons why the citizen has been flagged to verify that it is not a false positive, (4) if it turns out to be a false positive, the human reviewer restrains the system from flagging that decision subject again.

When the system is not developed in-house and responsibility is distributed across actors (e.g., dataset creators, model developers, system consumers), P13 pointed to the complexity of attributing responsibility correctly.

> *"If the City Hall outsourced the implementation of the system, then the outsourcing company would be responsible for correcting the system. But the citizen is unaware of that aspect and vis-a-vis the citizen the ultimate responsible is the City Hall. Then, ultimately, the City Hall should take responsibility"* (P13)

Certifying the algorithmic system before deployment was suggested as a means of unburdening the decision subject and ensuring a fair responsibility attribution.

T3.2. Fostering Social Transparency

Throughout the interviews, we observed that the unique nature of the public administration (e.g., far-reaching impacts, goals of social good) shaped the way in which our participants reacted to the presented decision-making process. On the one hand, because of the nature of public administration, a few participants requested transparency of cooperative activities (e.g., how AI implementation projects take place) in the context of the public administration (i.e., there were requests for *social transparency* [126, 383]). Social transparency within public administration was seen as a pre-requisite for implementing ex-ante contestability mechanisms. This was translated, for instance, in requests for participatory development approaches. To avoid corrective measures, P1 highlighted the importance of *probationary periods*. Probationary periods should be conducted in a way that does not impact ongoing activities and should be used to issue first warnings. P11 suggested that the municipality should consult decision subjects around their preferences towards the system at the early stages of AI development.

> *"There is probably a research team that has time and resources to organize 30-min video calls with each case to have a discussion like this in the early stages. Where they show these slides, and they have the different model architectures and data sources, potential harms, performance. Then I would be more interested."* (P11)

On the other hand, the nature of public administration led some participants to believe that the choices made during the system development were the correct ones. For example, P21 claimed that *"I really assume that they are indeed taking care that the data is good quality."* Similarly, because the public administration was the entity behind this system, some assumed that there would be more accountability and diligence when dealing with false positives. A few participants also made comparisons between the public and private spheres. Algorithmic decision-making processes in the public sphere were believed to be more contestable and were considered to have higher ethical standards.

3.5. Discussion

Our study aimed to generate in-depth *empirical* insights into *decision subjects'* information and procedural needs for meaningful contestability in a *high-risk* decision-making scenario in the public sector (i.e., an illegal holiday rental detection scenario). To this end, we conducted 21 interviews with participants with experience renting their properties out with varying levels of AI literacy. Instead of conceiving their right to autonomy as purely individual self-determination, our results suggest that participants' capacity for contestability was shaped and dependent on their interactions with other actors involved in decision-making. In this section, we summarize our results and position them in existing literature. We then discuss the implications for practice and research of our work.

3.5.1. Results in Relation to Previous Work

Information Needs for Contestability. Our results show that decision subjects have different strategies for deciding which information to request when contesting an algorithmic decision. These strategies depend on the perceived actionability of the provided information, and the risks and benefits of contesting the decision-making process. Regardless of AI literacy, there is a consensus in prioritizing the *why* (i.e., reasons, proof [179, 355]) behind the decision as a first step towards exercising their right to contest. The extent to which decision subjects want to know *how* the decision-making process took place depends on their AI literacy. It also depends on participants' experience with AI fairness. Especially among subjects with low AI literacy, knowing *how* the decision was made is not a priority. Among those who are interested in knowing *how* the decision was made, and unlike previous work on human-AI collaboration [219] and recommender systems [370], decision subjects do not want to "know everything". They are rather selective in choosing relevant information about the system that could help them contest the decision-making process [8]. This could be due to the differences in purpose (i.e., the aim of contesting *vs.* improving human-AI collaboration *vs.* getting better recommendations) and our participants' intrinsic need for practically helpful information because they are hypothesizing around a contestation scenario. The object and means of contestation (*what* participants in our study want to contest and *how* they want to proceed), in turn, depend on the perceived appropriateness of the information they receive and their ability to understand and use it as part of their contestation. Our findings further suggest that the sense-making process that preceeds a contestation is a cooperative process that participants engage in through expert advice or through dialogue with controllers. The means that enable such dialogue (i.e., communication channel, explanation medium), therefore, also affect the sense-making process. Even if theoretical claims have recognized the importance of justifications [179], or explanations [269] for contestability, there has been a comparatively small emphasis on empirically examining how decision subjects (individually or collectively) make sense of that information and how this empowers them to contest an algorithmic decision.

Procedural Needs for Contestability. Participants in our study request support from the decision-making organization and from peers to deal with the contestation process. This includes the presence of a third party to balance power and knowledge differentials. This suggests that participants perceive how algorithmic systems widen power gaps because of their complexity and opacity [269]. Our results further show that, for contesting aspects of the decision-making process that involve the algorithmic system itself, participants deem *collective action* as more effective than individual appeals. Even if the possibility of collective action was tangentially mentioned in theoretical frameworks [9], the insights from our participants provide detailed descriptions on what the collective could look like (e.g., led by AI experts) and what would define collective success (e.g., media attention, turning the issues into a political matter). When dealing with algorithmic failures (e.g., false positives), our results suggest that individual decision subjects do not want to bear the burden of identifying and contesting such failures. Participants in our study suggest that algorithmic failures should be corrected by *human controllers* (i.e., street-level bureaucrats involved in the first-order decision-making [9]). The very act of having to go through the process of contesting a false positive is considered to be unfair. Our participants also request transparency of the cooperative work that happens among *actors at previous stages of the AI development and deployment pipeline* as well as due responsibility attribution. The need to ensure transparency and due responsibility in a chain of distributed actors is aligned with theoretical claims for contestability by design [9]. It highlights the need to ensure awareness of risks and responsibilities across decision chains.

3.5.2. Implications for Practice

This section highlights the implications that our work has for public agencies integrating AI systems in decision-making processes.

Building Capacity for Supporting Contestability. The needs we identified for enabling decision subjects to meaningfully engage in acts of contestation are in tension with the challenges for contestability found by Alfrink et al. [7]. Those challenges include limited capacities of civil servants, organizational limits or resource constraints. These tensions indicate that there might be a mismatch between the capacity required to *ideally* address decision subjects' needs during contestation processes and the *reality* of what public administration can offer them in practice based on the available resources. For decision subjects to feel heard and understood, a balance between decision subjects' needs and the allocation of limited resources needs to be found. While participants in our study were generally not against using algorithmic systems for first-order decision-making (this could lead to public savings), they did insist on having a *human* reviewer during the contestation process. However, the organizational challenges of redistributing resources (e.g., economic, human, infrastructural) from the first-order decision-making to the contestation loop cannot be ignored. This is especially true when the algorithmic system suffers from functionality failures [333] in a resource-deficient context [356]. First, it is important to consider *who* is involved in the first-order decision-making and *who* in the contestation loop. How actors involved in different phases of the process have access to each other's information [92], the extent to which there is effective communication between them [333] or the scrutability of the system that mediates the process [269] are all aspects that make organizational change challenging. Furthermore, the relationship between the resources allocated for the current first-order human-led decision-making process and the resources needed for future contestation processes might not be a one-to-one relation. If such algorithms malfunction [333] and human oversight is motivated by legal compliance rather than quality control [7, 160], the harms generated when deployed at scale might multiply. It is, therefore, important to first ensure effective human oversight through e.g., explanations, cognitive forcing functions, or reinforcement learning paradigms [72, 73, 411]. Once appropriate human oversight is ensured, one way to build capacity for contestability would be to partly augment human reviewers' capacities (e.g., through chatbots [268] or methods to detect insincere contestations [27]) while ensuring decision subjects *feel heard*. If human oversight mechanisms are not effective or the option to augment human reviewers' capacities does not allow decision subjects to be heard and to exercise their right to contest automated decisions meaningfully (Article 22(3) of the GDPR [416]), the usage of AI systems might need to be interrupted.

Enabling Collective Action. Our results suggest that participants sought organizational and peer support to engage in acts of contestation. The conception of contestability might, therefore, need to account for the social nature of contestability. One way to do so is through collective contestations. Designing for collective contestability can involve indirect forms of control [7] through *representative bodies* of decision subjects [93, 407]. Examples of collective contestations include the *Contestation Café* suggested by

Collins and Redström [93]. The *Contestation Café* [93] is a speculative concept for community contestation, where decision subjects could learn to identify and contest unfair decisions. In a similar vein, *end-user driven audits* [108, 112, 238, 372] use the lived experiences of everyday users of algorithmic systems to uncover harmful algorithmic behaviors, which has, in turn, led to collective contestations (see Shen et al. [372] for a list of examples). An alternative line of work rather explores collective contestations as ex-ante mechanisms by e.g., involving decision subjects in the early stages of the AI design pipeline. This allows decision subjects to get actively involved in crafting the desired algorithmic behavior and in avoiding harmful consequences downstream [9]. If public agencies decide to explore this option, participatory frameworks such as *WeBuildAI* [250] could represent an interesting starting point. In *WeBuildAI* [250], stakeholders– including decision subjects–, can represent their views through computational models that contribute to algorithmic policy creation. For collective action like *Contestation Cafés* [93], end-user-driven audits [372] or WeBuildAI [250] to be of any use, participation is required. Participation, in turn, requires incentives [108] (e.g., available time, interest). An option to promote collectives could be for public administration to (financially) sustain them while ensuring collective action remains independent from the decision-making entity.

Defining Normative Boundaries for Contestability. A number of policy decisions should precede these contestation acts. These include determining what can be contested (both ex-ante and post-hoc), who can contest algorithmic decision-making processes, who is accountable for them, and what type of reviews or scrutiny mechanisms should be put in place [269]. Our work urges policymakers to further define normative boundaries for contestability.

3.5.3. Implications for Research

This section elaborates on the implications of our work for the CSCW research community.

Characterizing Individual and Collective Sense-Making of Personalized Explanations. According to our results, for a piece of information to be *actionable*, this information needs to be relevant and translatable into an "effective goal-oriented action" [408] (i.e., contesting). The relevance and potential of an information item to be translated into action, in turn, depends on decision subjects' ability to make sense and critically reflect on it to evaluate its appropriateness [362]. Our results, therefore, suggest that personalized, actionable explanations might be needed to address decision subjects' varying information needs for contestability. In contrast to *actionability* in *recourse* (i.e., set of factors that can be changed to obtain the desired outcome [214, 258, 379]), when dealing with contestability that goes beyond the decision outcome (i.e., it concerns the whole lifecycle of the system [9]), there is not one single definition for *actionable information.* There is, therefore, not a single response as to what information empowers decision subjects to meaningfully contest an algorithmic decision-making process [355, 378]. For explanations to be actionable for different decision subjects, they should, therefore, afford varying levels of sufficiency (i.e., content depth) and configuration. This could be operationalized by, for example, implementing explanations with hierarchies of information and varying levels of detail [101] or making explanations interactive [235]. Furthermore, the sense-making process of those explanations is not necessarily an individual process. It is additionally influenced by the actors that decision subjects could contact for help (e.g., legal representatives, human controllers). Further research is needed to know how different decision subjects make sense –individually and collectively– of personalized actionable explanations that are aimed explicitly at enacting contestability and that present varying levels of (1) availability, (2) content, (3) detail, (4) modality (i.e., audio vs. visual), and (5) paradigm (i.e., textual vs. graphical vs. interactive) [431]. Previous work on personalized explanations for recommender systems [394] could represent a good starting point for exploring personalized, actionable explanations *for contestability*. Personalized explanations for contestability will have to navigate the tension between opening algorithmic systems to scrutiny, and the need to align with privacy and confidentiality requirements [438].

Opening Up Sites for Contestation in AI Development and Deployment Pipelines. For contestability to be exercised by stakeholders (other than decision subjects) at earlier stages of the AI pipeline, tools that enable "real-time questioning, curiosity, and scrutiny" [226] of algorithmic systems by human controllers are needed. While some tools are already available that enable the scrutiny of algorithmic systems to surface information about decisions and models (e.g., What-if Tool [423]), further research is necessary to identify the needs of professional human controllers to interactively shape algorithmic behaviour and prevent false positives from repeatedly happening [226]. For due responsibility attribution across *algorithmic supply chains,* [92] tracking and documenting data flows represents the first step towards contestability—documentation which is required by the EU AI Act [133]. Exercising contestability throughout the algorithmic supply chain could, in turn, represent a step towards a deeper engagement with the system [405]. It would help actors distributed across the supply chain not only gain visibility over the supply chain itself, but it would also allow those actors to be attributed due responsibility when required. Enabling contestability throughout algorithmic supply chains [92] faces two main challenges that would benefit from further research. First, documenting discretionary choices made throughout the development and deployment pipeline of algorithmic systems is not straightforward [436]. There is a need to raise awareness around the value-laden (and therefore contestable) nature of "undisclosed yet impactful" [78] choices made throughout the pipeline. There is also a need to provide resources for practitioners to identify and effectively document such choices [36, 175, 274]. We echo prior work [184, 436] and encourage the CSCW community to look into strategies for scaffolding collaborative reflexive practices throughout the AI development and deployment pipelines. Second, even if those choices are acknowledged and documented, different actors across the supply chain might suffer from accountability horizon (i.e., limited capacity for system designers to understand the deployment context and for system consumers to influence its design) [92]. Therefore, legal and institutional mechanisms would be required to ensure visibility and influence over those design choices [92].

3.6. Caveats and Limitations

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

Participant Recruitment. To answer our research question, we sought to generate an in-depth understanding of decision subjects' needs for meaningful contestability and, therefore, decided to conduct qualitative interviews. In line with the Big Q qualitative research paradigm [90], we used purposive sampling to recruit participants that could help us generate nuanced insights into those needs. We, therefore, ensured that we had a diverse pool of participants in terms of AI literacy and ensured that the number of participants with low, medium, and high AI literacy was equally distributed. Among our participants, there was a more prominent representation of two countries (i.e., Netherlands and Spain). Similarly, our interviewees were all highly educated individuals (i.e., all had at least a bachelor's degree) and were used to interacting with digital platforms. Even if these choices are an intrinsic trade-off of Big Q qualitative research [90] in favor of generating in-depth insights, we acknowledge that our study might be subject to *representativeness limitations* [252].

Material Used for the Interviews and Transferability of Results . The letter we used as a prompt in the interview was designed and informed (e.g., penalty, contestation timeframe) by the guidelines that the Dutch public administration follows. Such a choice was made due to the origin of the suggested AI system and its specifications (i.e., the municipality of Amsterdam). The interviews, however, did not necessarily include citizens dealing with the Dutch public administration. Only a few participants mentioned the discrepancies between their experience with public administration communications and the material we presented. They considered this to be an irrelevant detail (e.g., *"They give me a timeframe. I don't care if it's 30 days [contestation timeframe in their residence country] or 6 weeks or whatever."* (P10)). However, we acknowledge that this mismatch might have affected how some other participants engaged in the interview. Similarly, the materials used for the interviews were based on a single case: a risk-scoring scenario for fraud detection within the public sector. We expect our findings to be transferable to other contexts where AI systems are used as part of policy enforcement efforts in the public sector. The transferability of our results to contexts other than policy enforcement support in the public sector will need further verification and should not be fully assumed.

Reflections on External Validity. For exploring the usage of algorithmic systems that have not yet been deployed, previous work has shown that scenario- or vignette-based qualitative methods can be useful instruments [70]. Several studies have also shown that how people react to studies in a "lab-based" environment is a good approximation to how they would react in the real world [427]. Furthermore, our recruitment strategy (i.e., participants who have experience renting their homes out) ensured that our participants had a hypothetical personal stake in the topic, as suggested by Clarke and Braun [90] when using interviews for capturing people's perceptions and understandings about a specific topic. However, a few of our participants indicated that they would not take the time to look at the information sheet (see Section 3.3.3 for information about the materials we used) if they had not been required to do so as part of the interview. In some cases, it was when participants engaged with the information sheet that they were able to raise concerns about the algorithmic system. This affected the object of contestation (i.e., *what* they wanted to contest). Results might have varied if participants were to contest a real-world algorithmic decision and had not inspected the information sheet.

3.7. Chapter Takeaways

In this chapter, we provided in-depth and nuanced empirical insights into the operationalization of contestability in algorithmic decision-making processes based on decision subjects' information and procedural needs. Through qualitative interviews, we found that decision subjects cooperation during the sense-making process that enables contestability. Strategies that participants used for making sense of the provided information varied based on participants' AI literacy and experience with AI fairness. Decision subjects additionally asked for support mechanisms both from the decision-making organization and from fellow decision subjects to effectively engage in acts of contestation. Lastly, decision subjects requested ex-ante interventions to ensure accountability in algorithmic decision-making. The findings presented in this chapter suggest that making algorithmic decision-making processes contestable by design is far from a trivial transition from currently available appeal mechanisms for human-led decision-making. Traditional appeal mechanisms place considerable burden on individual decision subjects. Our findings in this chapter highlight the need to account for the cooperative work behind meaningful contestability. In chapters 4 and 5, we narrow down contestability to the elements that constitute the *right to contest* automated decisions as interpreted from Article 22(3) of the European Union's General Data Protection (GDPR) [131]. We then assess the effect of contestability on decision subjects' fairness perceptions. In chapter 6, we reflect on future research directions that deal with cooperation for contestability.

4

Decision Subjects' Fairness Perceptions Towards The Right to Contest Automated Decisions

In this chapter, we generate empirical insights into decision subjects' fairness perceptions towards contestability as interpreted from Article 22(3) of the General Data Protection Regulation (GDPR) [131] (**RQ3**). Decision subjects' right to contest automated decisions requires the presence of explanations, and appeal mechanisms, and is conditioned by the implemented safeguards, e.g., the presence of human oversight during decision-making. In this chapter, we study the individual *and* combined effects of these three elements on decision subjects' fairness perceptions. To this end, we conduct a user study ($N = 267$) investigating the effects of explanations, human oversight, and appeal mechanisms on informational and procedural fairness perceptions for high- and low-stakes decisions in a loan approval scenario. We find that the presence of explanations and appeal mechanisms contribute to informational and procedural fairness perceptions, respectively, but we find no evidence for an effect of human oversight. These results further show that both informational and procedural fairness perceptions contribute positively to overall fairness perceptions but we do not find an interaction effect between them. A qualitative analysis exposes tensions between information overload and understanding, human involvement and timely decision-making, and accounting for personal circumstances while maintaining procedural consistency.

This chapter is based on the following publication: **Mireia Yurrita**, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, Alessandro Bozzon. "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* (CHI '23).

4.1. Introduction

Motivated by concerns about bias and discrimination in algorithmic decision-making [314], recent work has developed fairness-aware algorithmic systems [18, 124, 439] that ensure outcome distribution equity [121, 170]. However, even when a decision-making process is fair by some objective standard, decision subjects might not *perceive* it as fair [248] if aspects such as the inscrutability and unaccountability often surrounding algorithmic systems [59] go against their standards of justice [53, 247, 316].[1] Perceptions of unfairness could, in turn, jeopardize end users' trust in normatively fair algorithmic decision-making processes and, therefore, be an obstacle for their broader acceptance [53, 119, 247, 316, 419]. That is why a growing body of human-computer interaction (HCI) literature now focuses on determining which factors – e.g., information cues [257] such as explanations [59, 116, 304, 365] and system attributes [257] such as human oversight[2] [114, 262, 264, 303, 419] or contestability [271, 406] – effectively contribute to decision subjects' fairness perceptions.

Despite making important contributions, previous HCI research investigating fairness perceptions in algorithmic decision-making has faced two important limitations. First, earlier work has largely studied information cues and system attributes in isolation (e.g., [271, 406]). Such an approach fails to consider the entangled nature of these cues and attributes and does not align with the scenarios contemplated by regulatory efforts such as the European Union's *General Data Protection Regulation* (GDPR) [416]. For example, decision subjects can only meaningfully exercise their *right to contest* an algorithmic decision when they have solid arguments, which require explanations of the decision-making process [355, 416]. Contestation mechanisms and explanations thus co-shape the procedural justice principle of correctability [253] and may, therefore, co-mediate decision subjects' perceptions of procedural fairness [162, 253]. Not considering these entanglements could lead to blind spots regarding how different factors that are theoretically claimed to affect fairness perceptions (e.g., [406]) actually contribute to these perceptions.

Second, prior work has mainly used one-dimensional approaches for measuring fairness perceptions [23, 116, 247, 271, 304, 317, 357, 419, 442]. Although measuring such *overall fairness perceptions* is useful for capturing a global perception of appropriateness [94], prior work on legal and organizational psychology has often advocated for capturing fairness perceptions across up to four different dimensions (i.e., *faceted fairness perceptions*) [82, 95]. These dimensions include perceptions towards the equitable allocation of outcomes (i.e., *distributive fairness perceptions*) [2, 111], the nature of the process that leads to those decisions (i.e., *procedural fairness perceptions*) [253, 259, 393] as well as the information (i.e., *informational fairness perceptions*) [56, 162, 371]

[1] According to Cropanzano et al. [102], *justice* is a multi-dimensional construct that studies *fairness* perceptions across each of its dimensions. For instance, *procedural justice* refers to a justice dimension that aims to capture fairness perceptions regarding the *process* of a decision (i.e., *procedural fairness perceptions*). Colquitt [94] refer to *faceted fairness* as measurements of appropriateness that evoke different justice dimensions.

[2] Throughout this chapter, *human oversight* refers to a configuration where human intelligence is applied to identify and correct potential mistakes made by an algorithmic system [14]. We also call this configuration a *hybrid* human-artificial intelligence (AI) decision-making process.

and the treatment (i.e., *interpersonal fairness perceptions*) [56] received by decision subjects. Capturing how dimension-specific fairness perceptions manifest may help identify problematic aspects of algorithmic configurations. Additionally, learning how these dimension-specific fairness perceptions combine could then inform the prediction of global perceptions of appropriateness [94]. We argue that prioritizing the measurement of overall fairness might impede the development of a nuanced understanding of how different factors contribute to different facets of decision subjects' fairness perceptions [95].

This chapter takes a first step towards a nuanced understanding of how different information cues (i.e., explanations) and system attributes (i.e., human oversight and contestability) co-mediate multi-dimensional (i.e., informational and procedural) perceptions of fairness. Given the task-dependent nature of fairness perceptions [23, 53, 59, 247, 303, 381], we account for the stakes of the task as an additional contextual factor. Three research questions guide our work:

- **RQ3.1:** Do explanations, human oversight, and contestability affect perceived informational and procedural fairness in algorithmic decision-making processes?
- **RQ3.2:** Do the stakes (high/low) involved in the decision have an effect on perceived informational and procedural fairness?
- **RQ3.3:** Do decision subjects' perceived informational and procedural fairness predict overall perceived fairness?

To address these research questions, we first conducted a preliminary study to surface the interplay between explanations, human oversight and contestability (Section 4.4.1). We then used these findings to design an online, preregistered[3] user study where participants were shown a fictional loan approval process (Section 4.4.2). The descriptions shown to participants included information about the decision-making process with or without explanations, with or without human oversight and with or without the right to contest the decision (**RQ3.1**). Each participant was randomly assigned to a low-stakes[4] (holiday) or to a high-stakes (home) loan approval scenario (**RQ3.2**). For each scenario, we measured perceptions of informational, procedural and overall fairness (**RQ3.3**).

Our results show that explanations and contestability affect decision subjects' informational[5] and procedural fairness perceptions, respectively (**RQ3.1**; see Section 4.5.2). We do not find evidence that decision subjects' perceptions of informational and procedural fairness are influenced by human oversight (**RQ3.1**) or the stakes of the task (**RQ3.2**). Our results further show that perceptions of informational and procedural fairness both relate positively to perceptions of overall fairness, but we do not find an interaction effect between them (**RQ3.3**). As part of our exploratory analyses, we unpack

[3] The preregistration is openly available at https://osf.io/4uf3m.

[4] Loan approval decisions are generally seen as high-stakes [98] but we still expect differences in users' perceived stakes depending on the loan purpose.

[5] This result replicates and confirms a finding from earlier work [365].

informational and procedural fairness perceptions into the sub-elements that compose each dimension (Section 4.5.3). We find that decision subjects may rate perceptions of procedural voice and outcome influence negatively, even when contestability (in the form of appeal processes) is incorporated. We also find that including human oversight may deteriorate perceptions of process consistency and lack of bias. Through a qualitative analysis, we identify three areas of tension: (1) amount of information vs. generating understanding for all, (2) human involvement vs. timely decision-making, and (3) standardized fact-based process vs. accounting for personal circumstances (see Section 4.5.4). These insights set the grounds for motivating the exploration of transparency beyond outcome explanations, for crafting alternative human-AI configurations, and for designing contestation mechanisms that effectively give voice to decision subjects.

Supplementary materials linked to this chapter include task design, preregistration, data, and code for statistical analysis and are openly available at https://doi.org/10.4121/62a7ad5f-1225-4618-bd4b-1d66a3941db3.

4.2. Related Work

This section describes previous research on how explanations, human oversight, and contestability contribute to fairness perceptions in algorithmic decision-making and discusses the task-dependent nature of this work. We focus on these specific information cues and system attributes as they are directly addressed by Article 22(3) of the GDPR [416]. We then cover research on human decision-making, where fairness perceptions have been captured across multiple dimensions.

4.2.1. Factors Affecting Perceptions of Fairness in Algorithmic Decision-Making

Explanations. Explanations (i.e., representations of a system's ability to account for their own operation in ways that help users understand how these tasks are being accomplished [59]) are considered key elements for enhancing users' fairness perceptions in algorithmic decision-making processes. Previous work has demonstrated the positive effect of different explanation styles on decision subjects' feelings of justice [59, 116] and their confidence in the fairness of algorithmic systems [316]. Schoeffer et al. [365] found that the amount of information in explanations was positively related to *informational* fairness perceptions.

Human Oversight. The term *human oversight* has been used to refer to the configuration where human intelligence is applied to identify potential mistakes in algorithmic decision-making processes [14]. Since algorithmic systems can perform increasingly complex tasks [422], recent research has pointed to opportunities for crafting more reliable and timely decision-making processes with human-artificial intelligence (AI) collaborations [39, 440]. Despite this growing interest, most recent work on fairness perceptions has focused on comparing AI systems with their human counterparts [23, 80, 114, 148, 233, 247, 262, 317] rather than comparing fully automated with hybrid configurations. In one study that did compare algorithmic decision-making to hybrid and human decision-making, Nagtegaal et al. [303] found that hybrid configurations can increase public employees' (subjects of managerial decisions) perceptions of procedural fairness. Wang et al. [419] also evaluated the effect of hybrid decision-making processes on decision subjects' perceptions of fairness but did not find any evidence that hybrid decision-making processes are perceived to be fairer than fully automated ones.

Contestability. Contesting a decision has been defined as the act of opposing an action; either because the action is perceived as mistaken or simply wrong [10, 412]. *Contestability* has, thus, been conceptualized as recourse [206, 404, 410], appeal [412], and as a design principle (i.e., *contestability by design*) [9, 14, 355]. Contestability is said to "surface values" [405] and to be a "form of procedural justice, a way of giving voice to decision subjects, which increases perceptions of fairness" [9]. To the best of our knowledge, however, the effect of contestability in algorithmic decision-making has not yet been widely studied. In one of the few studies that empirically tested the effect of appeals on decision subjects' perceptions of fairness, Vaccaro et al. [406] found that none of their appeal designs improved these perceptions.

Task stakes. Perceptions towards algorithmic decision-making can vary across scenarios [53, 59], based on task characteristics [247], and the stakes of the task (i.e., the impact that a negative outcome would have on the future of an individual [211]) [23, 303, 381]. For instance, Binns et al. [59] found that scenario effects obscure explanation effects under repeated exposure of one explanation style. Lee [247] saw differences in fairness perceptions towards human and algorithmic decision-makers based on task characteristics. Araujo et al. [23] argued that decision subjects may perceive algorithmic systems as fairer than human experts only for high-impact decisions in the justice and health domains.

4.2.2. Capturing Perceptions of Fairness in Decision-Making Processes

Decision subjects' perceptions of fairness can be complicated and nuanced [419]. To measure these perceptions in a granular way, disciplines in social sciences such as legal and organizational psychology have empirically validated models that capture perceptions of fairness across different dimensions [94, 102]. These dimensions include perceptions of fairness towards decision outcomes (i.e., *distributive fairness perceptions*) [2, 111], the processes that led to those outcomes (i.e., *procedural fairness perceptions*) [253, 259, 393], the treatment received by decision subjects (i.e., *interpersonal fairness perceptions*) [56], and the information given to decision subjects (i.e., *informational fairness perceptions*) [56, 162, 371]. Each of these dimensions evokes different justice principles and is built upon criteria that have been found to be relevant for that dimension [409]. For instance, procedural fairness perceptions are measured considering perceptions of *procedural voice*, *outcome control*, *consistency of procedures across participants*, *suppression of bias*, *accuracy of factors*, *correctability of outcomes*, and *ethicality of the process* [253, 393].

4.2.3. Research Gap and Motivation

Although earlier work has shed some light on how to go from a normative to a behavioral understanding of fairness, evidence on how factors that are theoretically related to certain principles of justice co-mediate decision subjects' perceptions of fairness in algorithmic decision-making is still lacking. One reason for this is that the effects of factors believed to enhance perceptions of fairness have been obscured by phenomena such as the *outcome favorability bias* (i.e., divergence in decision subjects' perceived fairness based on the favorability of the outcome they receive personally) [317, 419]. For example, although including human oversight has been claimed to bring together the best of the manual and the automatic worlds, there is still little insight into how human oversight contributes to decision subjects' perceptions of fairness. Similarly, although contestability has been claimed to be a key aspect to enhance perceptions of fairness, to the best of our knowledge, there is currently no empirical evidence on whether or how contestability contributes to these perceptions. One could argue that Lyons et al. [269] looked into different modalities of appeal processes and evaluated perceptions of fairness in each case. However, evaluating perceptions of fairness towards different types of appeals is different from evaluating perceptions of fairness towards an algorithmic decision-making process that offers the right to appeal. Another key limitation of previous research is that it did not consider the entangled nature of explanations, human oversight, and contestability. Although decision subjects' right to explanation is not explicitly guaranteed by the GDPR [368], Article 22(3) does explicitly guarantee their right to contest a negative decision [416], for which decision subjects need meaningful (i.e., functional [368]) explanations [355]. The GDPR also states that contestations might vary based on the human intervention in the original decision [416]. Therefore, the way in which a decision can be meaningfully contested depends on the received explanations [355] as well as the interpretation of the implemented safeguards (i.e., right to human intervention, right to express views, and right to contest the decision) [416].

From a methodological perspective, a majority of previous studies has used mono-dimensional (i.e., overall fairness perceptions [94]) approaches for capturing the effects of explanations, human oversight, and contestability on fairness perceptions [23, 116, 247, 271, 304, 317, 357, 419, 442]. This has resulted in a lack of nuance in the understanding of how fairness perceptions are co-mediated by each of these factors. We echo the need to include lessons from the replication crisis within psychology [66] and advocate for a multi-dimensional approach to measuring perceptions of fairness (i.e., faceted fairness perceptions [94]). Although these dimensions were suggested for human decision-making, we argue that they represent a good starting point toward developing standardized methods for specifically evaluating algorithmic decision-making processes. The benefits of using a more nuanced approach for measuring the effect of explanations on perceptions of fairness have already become evident. Schoeffer et al. [365] found that outcome explanations would increase decision subjects' perceptions of *informational* fairness, but it would make them question structural aspects of the *procedure*, just as it was claimed by Greenberg [162] for human decision-making.

In this chapter, we address the above gaps by systematically evaluating algorithmic decision-making processes with varying levels of explanations, human oversight, and contestability, and unpack and disentangle their effects on perceptions of fairness through a multi-dimensional approach. Since the factors (i.e., explanations, human oversight, contestability) that we manipulate in our experimental setting have been related to perceptions of informational and procedural fairness in human decision-making

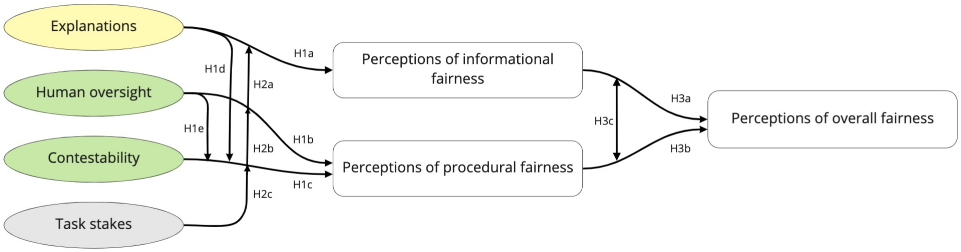

Explanations
Human oversight
Contestability
Task stakes
H1a
H1b
H1c
H1d
H1e
H2a
H2b
H2c
Perceptions of informational fairness
Perceptions of procedural fairness
H3a
H3b
H3c
Perceptions of overall fairness

Figure 4.1: Overview of the hypotheses. Yellow refers to information cues, green to system attributes, and grey to contextual factors.

[253, 371], we capture perceptions of fairness across those two dimensions. We also test the predictive validity [95] for these multi-dimensional fairness perceptions on overall fairness perceptions. This enables us to compare the multi-dimensional approach with previously used mono-dimensional approaches.

4.3. Hypotheses

Drawing from literature in legal and organizational psychology for human decision-making [22, 41, 56, 57, 162, 163, 400] and studies on perceptions of fairness in AI systems [23, 53, 166, 247, 316, 357, 365, 406, 419], we formulated eleven hypotheses (Figure 4.1). Each hypothesis is related to one of the research questions outlined in Section 4.1 and is followed by a rationale. We preregistered all hypotheses before data collection.

4.3.1. Hypotheses related to RQ3.1: Explanations, Human Oversight, and Contestability

- **Hypothesis 3.1a** ($H_{3.1a}$). Decision subjects perceive algorithmic decision-making processes as more informationally fair when they are accompanied with explanations.

  Rationale. We extend Schoeffer et al. [365]'s study to evaluate the effect of explanations on informational fairness in both high-stakes and low-stakes decisions. We expect to replicate their findings in our own experimental setting.

- **Hypothesis 3.1b** ($H_{3.1b}$). Decision subjects perceive algorithmic decision-making processes as more procedurally fair when these processes are supplemented by human oversight rather than fully automated.

  Rationale. Previous studies have found that decision subjects consider human decisions to be fairer than fully-automated, algorithmic decisions; especially for practices that are highly complex and are perceived to require human skills [247, 303]. Although recent research has found contradictory evidence on whether decision subjects perceive *hybrid* decision-making as fairer than entirely algorithmic decision-making [303, 419], we do expect that human oversight will lead to increased *procedural fairness* perceptions among decision subjects in sensitive contexts (e.g., loan approval processes).

- **Hypothesis 3.1c** ($H_{3.1c}$). Decision subjects' procedural fairness perceptions differ based on the contestation procedure of an algorithmic decision-making process.

Rationale. We hypothesize that, as with human decision-making [393], contestation procedures in algorithmic decision-making processes affect perceived procedural fairness.

- **Hypothesis 3.1d** ($H_{3.1d}$). The effect of contestability on decision subjects' procedural fairness perceptions is moderated by the presence of explanations.

  Rationale. Schoeffer et al. [365] found that, although including more information in explanations led to an increased perception of informational fairness, the presence of explanations allowed decision subjects to question the way in which different factors were being used for decision-making. We thus hypothesize that, aside from a general effect of contestability on decision subjects' procedural fairness perception (see H_{1c}), the presence of explanations and contestability on the algorithmic decision *interact* in affecting decision subjects' perceived procedural fairness.

- **Hypothesis 3.1e** ($H_{3.1e}$). The effect of contestability on decision subjects' procedural fairness perceptions is moderated by the presence of human oversight.

  Rationale. Various studies have demonstrated decision subjects' concern for fully automated, highly complex decision-making processes [247, 303]. That is why we expect that configurations where decision subjects can contest an algorithmic decision lead to varying degrees of procedural fairness perceptions in decision subjects depending on whether the original decision was made by a fully-automated or hybrid system.

4.3.2. Hypothesis related to RQ3.2: Task stakes

- **Hypothesis 3.2a** ($H_{3.2a}$). The effect of explanations on decision subjects' informational fairness perceptions is moderated by the stakes of the task.

  Rationale. Binns et al. [59] found that the nature of the presented scenario moderates the effect of explanation types on fairness perceptions. In line with these findings, we hypothesize that, based on the nature of the task at stake (i.e., involving high or low stakes), decision subjects will be satisfied differently with the amount of information they received.

- **Hypothesis 3.2b** ($H_{3.2b}$). The effect of human oversight on decision subjects' procedural fairness perceptions is moderated by the stakes of the task.

  Rationale. Lee [247] demonstrated that fairness perceptions regarding the decision maker (i.e., a fully-automated system or a human) were moderated by task characteristics. Nagtegaal et al. [303] also found that the effect of involving humans on perceptions of procedural justice varied based on the complexity of the task. Despite the context being different (both these studies focused on managerial decisions) and our study considering fully-automated vs hybrid decision making, we hypothesize that the stakes of the task (i.e., involving high or low stakes) will similarly moderate the effect of human oversight on procedural fairness perceptions in our study.

- **Hypothesis 3.2c** ($H_{3.2c}$). The effect of contestability on decision subjects' procedural fairness perceptions is moderated by the stakes of the task.

Rationale. Previous work has suggested that perceptions of fairness regarding the decision-maker generally depend on the nature of the task [247]. We thus hypothesize that the stakes of the task (i.e., involving high or low stakes) also moderate the effect of contestability (e.g., when decision subjects are given the right to contest the decision-maker [271]) on decision subjects' procedural fairness perceptions.

4.3.3. Hypothesis related to RQ3.3: Overall vs. Faceted fairness

- **Hypothesis 3.3a** ($H_{3.3a}$). Decision subjects' informational fairness perceptions are positively associated with their overall fairness perceptions.

  Rationale. This hypothesis is in line with findings in human decision making, where informational fairness was claimed to influence perceptions of overall fairness [95, 164].

- **Hypothesis 3.3b** ($H_{3.3b}$). Decision subjects' procedural fairness perceptions are positively associated with their overall fairness perceptions.

  Rationale. Studies dealing with procedural fairness in human decision-making processes [164, 393] demonstrated that participants with a strong influence over the decision-making process were more likely to perceive a negative outcome as fair [203]. We hypothesize that for algorithmic decision-making processes, there will also be a positive relation between perceptions of procedural fairness and overall fairness.

- **Hypothesis 3.3c** ($H_{3.3c}$). Decision subjects' perceived informational and procedural fairness interact in predicting overall fairness.

  Rationale. Research in human decision-making has demonstrated that explanations provide the "information needed to evaluate structural aspects of decision-making" [162]. In line with these findings, we hypothesize that perceptions of overall fairness are not just dependent on both informational and procedural fairness, but that these two factors *interact* in predicting overall fairness perceptions.

4.4. Study Design

Because explanations, human oversight, and contestability are entangled by nature [416], we first conducted a preliminary study to craft an experimental setting that would surface the interplay between these factors (Section 4.4.1). In this exploratory study, we captured preferences towards different explanation styles and investigated what aspects participants would like to contest. We then combined these insights with previous literature to design our main user study in the context of a loan approval process (Section 4.4.2).

4.4.1. Preliminary Study

This preliminary study ($N = 58$) aimed at crafting (1) understandable and (2) actionable[6] explanations that (3) support contestability [416]. We also sought to understand what

[6] We define "actionable" factors as the set of variables upon which interventions are possible. We include those variables that may change as a consequence of a change to its causal ancestors (that other authors have named as "mutable but non-actionable" [214])

aspects of the decision-making process participants may contest. Although prior work has already studied the understandability of different types of explanations [59, 116] and identified actionable factors for loan approval processes [365], the interplay between explanations and contestability still represents an underexplored area,[7] hence the need to perform this preliminary study. The design of our preliminary study and the instruments we used to capture participants' preferences can be found in our repository.

Method of the Preliminary Study.

As part of our preliminary study, we provided each participant with five types of explanations (randomized) for a fictional home loan denial scenario: (1) *factor importance-based* explanations (i.e., feature importance hierarchy using ">" for expressing "more important than" [365]), (2) *input influence-based* [8] explanations (i.e., list of input variables along with a quantitative measure of the effect and directionality —positive or negative— that each of these variable had on the final decision [59, 116]), (3) *case-based* explanations (i.e., instance from the model's training data that is most similar to the decision being explained [59, 116]), (4) *counterfactual* explanations (i.e., representation of the alterations that input variables would need for the undesired model output to change [59, 116, 416]), and a combination of (5) *input influence-based and counterfactual* explanations [365]. They were then asked to select the two most understandable and actionable explanations and two explanations thanks to which the decision subject would best know what information to use to contest the decision. We also asked them to choose their overall preferred explanation type. At the end of the study, we included two open-ended questions. The first question aimed to disclose the rationales behind participants' preferences for different types of explanations. The second question collected answers on what aspects of the decision-making process participants would be willing to contest. For analyzing the responses to the open-ended questions, we performed a reflexive thematic analysis [68]. Our aim was to use the findings from this preliminary study to inform the design of our main user study (Section 4.4.2).

Insights from the Preliminary Study.

The combination of counterfactuals and input influence-based explanations scored highest for all criteria (see Table 4.1). To better understand these results, we discuss our findings from the qualitative analysis below. We refer to quotes as Q.*i*, where *i* is the index of a specific quote. Appendix B.1. shows all selected quotes.

Preferences towards different types of explanations. In line with findings from Dodge et al. [116], we found that case-based explanations were considered less fair (Q.1, Q.2). Participants generally preferred explanations that contain more information, which is in

[7] Although the interplay between explanations and recourse is increasingly being studied (e.g., [213, 384]), for this preliminary study, we do not limit contestability to recourse and inquire whether participants would question other aspects of the decision-making process.

[8] As opposed to some previous work [59, 116], where the quantitative measurement of the input influence was indicated through a varying number of "+" (positive influence) or "-" (negative influence) signs, we expressed this difference in influence through numerical values. We clarified that the number in brackets indicated the magnitude of the positive or negative effect that the variable had on the final decision —negative meaning a contribution towards the rejection decision—.

Explanation type	Understandable	Actionable	Supports contestability	Overall
Importance-based	23.64%	17.70%	18.35%	12.08%
Input influence-based	17.27%	20.35%	21.10%	15.52%
Case-based	16.36%	8.85%	14.68%	13.79%
Counterfactual	13.64%	15.93%	16.51%	15.52%
Counterfactual & input influence-based	**29.09%**	**37.17%**	**29.36%**	**43.10%**

Table 4.1: Results from our preliminary exploratory study. We evaluated how (1) understandable and (2) actionable different types of explanations were, and to what extent they (3) supported contestability. Column (4) shows participants' overall preferred option.

line with findings from Schoeffer et al. [365] (Q.3). Moreover, participants generally preferred the combination of input influence-based and counterfactual explanations because these included descriptions of the "how" and a justification of the "why" of decisions, as suggested by Sarra [355]. Input influence-based explanations were regarded as faithful descriptions of how each feature contributes to the algorithm's decision-making process (11/58)[9] (Q.4). Despite using numerical values to indicate different degrees of input influence on the final decision, readability was not flagged as an issue for input influence-based explanations by our participants. Counterfactuals were regarded as concise and explicit when directing the attention to features that were relevant to that particular decision (17/58) (Q.5, Q.6).

What to contest. Participants pointed to two main aspects they would like to contest: first, the basis (i.e., the factors) of the decision and their weights (28/58) (Q.7, Q.8) and second, the usage of an AI (10/58). AI systems were viewed as lacking subjective judgment capabilities for considering individual circumstances (in line with previous studies [80, 247, 303]) (Q.9). Generalization was also considered to be an inappropriate basis for decision-making (Q.10).

4.4.2. Main User Study

In our main user study, we sought to characterize the main and interaction effects of explanations, human oversight, and contestability on perceptions of informational and procedural fairness. We also explored the influence of contextual factors (i.e., the stakes of the task) in this context and captured the relationship between informational and procedural fairness perceptions and perceptions of overall fairness. We had preregistered our hypotheses, research design, and data analysis plan for the main study before data collection.

Independent Variables

In an effort to minimize the effect of *outcome favorability bias* [419], we followed prior research [23, 365, 381] and showed participants in our user study a fictional loan approval

[9] We indicate the prevalence of each statement using proportions (a/b), where *a* indicates the number of participants whose response to the open-ended questions was related to the statement in question, and *b* indicates either the number of participants within a condition that we are specifically referring to or the total number of participants in the study (58 for the preliminary study and 267 for the main study).

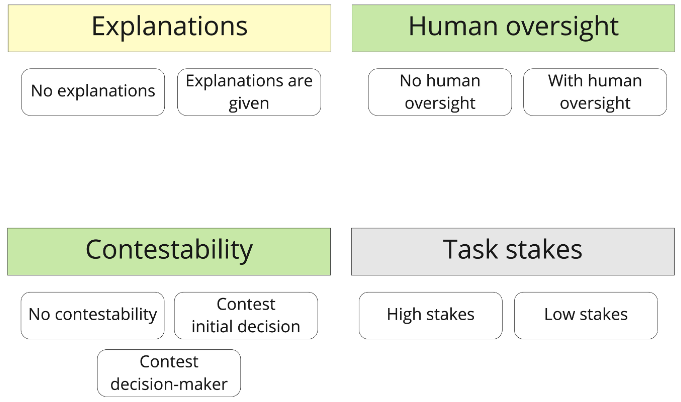


Figure 4.2: Overview of the independent variables. Yellow refers to information cues, green to system attributes, and grey to contextual factors. White colored boxes indicate the conditions we controlled for each factor.

scenario involving the fictional character *Kim* as loan requester. The scenario differed depending on four independent variables. Figure 4.2 gives an overview of the independent variables and Table 6.6 in Appendix B.2 shows how each independent variable was displayed in practice.

- *Explanations* (categorical, between-subjects). We assigned each participant to one of two configurations:

  1. No explanation: participants saw what information the fictional loan requester had been asked to provide but not how this information was used.
  2. With explanation: participants learned the weight each piece of information had in the final decision (*input influence-based explanation*) and the hypothetical scenarios where the loan requester would have been able to have the loan approved (*counterfactuals*). The factors requested by the bank and the given explanations are inspired by prior work [365] and enhanced based on the insights we got from the preliminary study (Section 4.4.1). We discarded gender and marital status as decision basis because these factors are explicitly protected by law [46]. Note that the *no explanation* configuration in our study is equivalent to the *disclosure of factors* condition defined by Schoeffer et al. [365], and not to the *baseline without further explanations*. The rationale behind this design choice is twofold: first, we argue that the disclosure of these factors is necessary for participants to be able to judge the fairness of the decision basis. Second, Schoeffer et al. [365] found no difference in informational fairness perceptions between the two aforementioned configurations. These explanations were textual to limit presentation complexity [53, 84, 365].

- *Human oversight* (categorical, between-subjects). We randomly assigned each participant to one of two configurations:

1. No human oversight: participants were told that the algorithmic decision-making process was fully automated.
2. With human oversight: participants were told that the loan approval process combined the usage of an algorithmic system with human expertise. We designed this condition based on one of the human-in-the-loop configurations discussed by Almada [14]. As opposed to some previous work where a human would supervise each decision made by the algorithmic system [419] — the authors did not find any evidence of this configuration affecting fairness perceptions—, in our study human intervention would serve as a quality control against machine failures [14]. We, therefore, used the confidence of the prediction as an indicator of a potential mistake made by the algorithmic system. The approval process would involve two steps: a first step where the algorithmic system receives an online loan request and evaluates the case; and a second step where a human expert [317] (bank employee) oversees the decision if the algorithmic decision-making system's confidence is low.

- *Contestability* (categorical, between-subjects). We designed contestation mechanisms in the form of appeal processes, following findings from our preliminary study (Section 4.4.1) and previous literature [271, 416]. Participants in our preliminary study mainly wanted to contest (1) the algorithmic decision-maker or (2) the basis of the decision. These strategies resonated with the *new information condition* and *new decision condition* (with a human reviewer) defined by Lyons et al. [271]. We randomly assigned each participant to one of three configurations:

  1. No contestability: participants were told that, due to time constraints, there would be no option for the fictional loan requester to contest the decision in case of a rejection.
  2. Option to contest the initial decision and provide additional information: participants were told that, in case of a rejection, the fictional loan requester had the option to make objections about the initial decision and provide any information to support the application. The same system (if a human oversaw the initial decision, the same human would oversee the review process) would reevaluate the loan application.
  3. Contest decision-maker: participants were told that, in case of a rejection, the fictional loan requester had the opportunity to ask a human (different from the one who oversaw the process if there was already a human involved in the initial decision) to review the process. This human reviewer would make a completely new decision with the information that Kim had already provided for the initial decision.

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

  1. High-stakes decision: the purpose of the loan application is to buy a house.
  2. Low-stakes decision: the purpose of the loan application is to go on a holiday trip.

Dependent Variables.
The instruments we used to measure the dependent variables can be found in our repository.

- *Perceptions of informational fairness* (continuous). Measured by the average score on four of the items used by Schoeffer et al. [365], based on Bies and Moag [56] and Greenberg [162].

- *Perceptions of procedural fairness* (continuous). Measured by the average score on the seven items defined by Colquitt [95],[10] based on Thibaut and Walker [393] and Leventhal [253].

- *Perceptions of overall fairness* (continuous). Measured by a single item rated on a seven-point Likert scale [239, 247].

Descriptive and Exploratory Measurements
The instruments we used to measure the descriptive and exploratory variables can be found in our repository.

- *Age group* (categorical). Participants selected their age group from multiple choices.

- *Level of education* (categorical). Participants selected their highest completed level of education from multiple choices.

- *AI literacy* (continuous). AI literacy has been proven to significantly affect perceptions of informational fairness [365]. We, therefore, captured the average score of the four items defined by Schoeffer et al. [365].

- *Affinity to technology* (continuous). Langer et al. [239] showed that affinity to technology was consistently correlated with end users' perceptions of algorithmic capabilities. We, therefore, captured the average score of the four items defined by Franke et al. [143] as a possible control variable.

- *Personal experience* (continuous). Kramer et al. [233] showed that preferences towards humans vs. algorithmic systems depend on people's previous experience with the described situation. We, therefore, captured the average score of the two items defined by Kramer et al. [233].

- *Task stakes perception* (continuous). Since the stakes involved in a decision are subjective and personal [211], we captured participants' task stakes perceptions as a manipulation check. This was measured through an adapted version of the item defined by Lyons et al. [271].

[10] After pilot testing the wording and layout of the presented scenarios, we rephrased some of the items to make them more understandable for participants.

Procedure

The study consisted of four main steps.

Step 1. Participants stated their age group and level of education. Their degrees of AI literacy, affinity to technology, personal experience and task stakes perception were also measured.

Step 2. Participants were presented with a fictional loan approval scenario involving a person named Kim. Previous research has shown that under *repeated interactions* with algorithmic decision-making systems, decision subjects' fairness perceptions are affected by the favorability of the system towards the group that the decision subjects belong to [154]. In order to minimize these effects, we limited our study to a *one-shot interaction* with the system and we did not disclose the demographics of Kim, such as their gender and age. Kim had applied for a loan online and was waiting for the bank to assess their eligibility. Depending on the stakes of the task that participants had been assigned to, the purpose of this loan would be either to buy a house (high stakes) or to go on a holiday trip (low stakes). Participants would be informed about the information Kim had provided to the bank to evaluate the loan request. As part of the scenario, every participant would then be informed that Kim's loan request had been rejected and they would get to know the process through which the loan request had been evaluated. Based on which of the $(2 \times 2 \times 3 \times 2) = 24$ between-subject scenarios a participant had randomly been placed in, participants would receive explanations about the outcome of the decision, learn whether there was a human expert overseeing the process and get information about whether and how Kim could contest the decision (see Table 4.2). Participants would then respond to an attention check, where they would be asked about the purpose of the loan request.

Step 3. Participants evaluated their perceptions of informational, procedural, and overall fairness. Additionally, this step included a second attention check that asked participants to select a specific option from a Likert scale.

Step 4. Participants were asked two optional open-ended questions to describe what kind of information they would have liked to receive (if any) and what element would have made the decision-making process fairer (if any).

Data Collection

We planned to collect data from at least 261 participants. We computed the required sample size using the software *G*Power* [136] for an ANOVA with main effects and interactions; specifying the default effect size of 0.25, a significance threshold of $\alpha = \frac{0.05}{11} = 0.0045$ (i.e., due to testing multiple hypotheses; see Section 4.4.2), a desired power of 0.8, 24 groups, and the respective degrees of freedom for the different hypotheses we aimed to test.

We recruited 279 participants from *Prolific* (https://prolific.co). Each participant was at least 18 years old, had high proficiency in English, and could participate in our study only once. Participants were rewarded based on a $12 hourly rate and the median completion time was 7 minutes and 41 seconds. Participants were excluded from data analysis if they did not pass at least one of the attention checks in the experiment. This led to a total number of 267 participants. The study itself was conducted on *Qualtrics* (https://www.qualtrics.com), where participants authenticated with a registration to-

A bank has implemented a new loan application system where potential customers apply for a loan online and then the company assesses the eligibility of the customer for the loan.

<Configuration [*No human oversight*] or [*With human oversight*]>

Kim, a potential customer, is looking for funding opportunities to <task> and has thus decided to apply for a <task> loan through the bank's online platform. As part of the <task> loan application process, the bank has requested the following information:

- Applicant annual income
- Co-applicant (if any) annual income
- Credit score
- Date of birth
- Employment status
- Education
- Loan amount requested
- Loan amount term (months)
- Loan purpose
- Number of dependents

A few hours after sending the requested information, Kim has received an email with the final decision: the loan has been rejected.

<Configuration [*No explanation*] or [*With explanations*]>

<Configuration [*No contestability*] or [*Contest initial decision*] or [*Contest decision-maker*]>

Table 4.2: Overview of the scenario.

ken received on *Prolific*. Our study was approved by a research ethics committee at our institution.

Statistical Analyses

Before conducting any statistical analyses, we mapped all (seven-point) Likert scale answers onto an ordinal scale ranging from −3 (i.e., strongly disagree) to 3 (i.e., strongly agree) and computed averages for answers on related items (e.g., to obtain participants' informational and procedural fairness perceptions).

We analyzed the hypotheses we specified in Section 4.3 in three separate statistical analyses. First, to test $\mathbf{H_{3.1a}}$ and $\mathbf{H_{3.2a}}$, we conducted a multi-way ANOVA with *explanations, human oversight, contestability*, and *task stakes* as between-subjects factors and *perceptions of informational fairness* as dependent variable.[11] Second, to test $\mathbf{H_{3.1b\text{-}e}}$ and $\mathbf{H_{3.2b\text{-}c}}$, we conducted another multi-way ANOVA with the same between-subjects factors but with *perceptions of procedural fairness* as the dependent variable. Third, to test $\mathbf{H_{3.3a\text{-}c}}$, we conducted a multiple linear regression analysis with *perceptions of informational fairness* and *perceptions of procedural fairness* as independent and *perceptions of overall fairness* as dependent variables. Because we were testing 11 hypotheses as part of this study, we applied a Bonferroni correction to our significance threshold, reducing it to $\frac{0.05}{11} = 0.0045$. This means that p-values resulting from the analyses described above are only regarded as significant if they are below this reduced threshold. Next to the F statistic and p-value, we also report the partial eta squared (η_p^2) effect size for each hypothesis test that was part of an ANOVA.

In addition to the analyses described above, we conducted posthoc tests (i.e., to analyze pairwise differences), Bayesian hypothesis tests[12] (i.e., to quantify evidence in favor of null hypotheses), and exploratory analyses (i.e., to note any unforeseen trends in the data) to better understand our results. We also performed a qualitative, reflexive thematic analysis [68]. The first author coded the responses to the open-ended questions inductively using *Atlas.ti* (https://atlasti.com). These codings were grouped into themes and iteratively refined.

4.5. Results

In this section, we analyze the results of the main user study (see Section 4.4.2). Table 4.3 shows a summary of our results.

4.5.1. Descriptive Statistics

Of the 267 participants in our user study, 19.5% were between 18 and 25 years old, 35% between 26 and 35 years old, 28.5% between 36 and 50 years old, and 17% were between 50-80. 60% of the participants had at least a Bachelor's degree. 87% of our participants claimed to have heard or had experience with humans making loan decisions, whereas

[11] Although we did not specifically hypothesize about the effects of human oversight and contestability on informational fairness perception, we included these variables here for exploratory analyses.

[12] Depending on the outcome of the relevant classical hypothesis test, we report Bayes Factors in favor of the alternative hypothesis (BF_{10}) or the null hypothesis (BF_{01}). We interpret the Bayes Factors according to the guide by Lee and Wagenmakers [243] who adapted it from Jeffreys [202].

72% of them had heard of or had experience with an algorithmic system making the decision.

4.5.2. Hypothesis Tests

Our first confirmatory analysis was a multi-way ANOVA with the presence of explanations, human oversight, contestability, and task stakes as between-subjects factors and perceptions of informational fairness as the dependent variable. We found a main effect of the presence of *explanations* ($\mathbf{H}_{3.1a}$; $F(1,260) = 74.21$, $p < 0.001$, $\eta_p^2 = 0.22$; $\mathrm{BF}_{10} > 1000$) on decision subjects' informational fairness perceptions. However, we did not find any evidence indicating that the effect of explanations on informational fairness is moderated by *task stakes* ($\mathbf{H}_{3.2a}$; $F(1,260) = 0.01$, $p = 0.92$, $\eta_p^2 < 0.01$). A Bayesian analysis revealed moderate evidence in favor of the null hypothesis that there is no such interaction effect ($\mathrm{BF}_{01} = 7.44$).

The second multi-way ANOVA analysis we conducted had the presence of explanations, human oversight, contestability, and task stakes as between-subjects factors and perceptions of procedural fairness as the dependent variable. We did not find any evidence of *human oversight* impacting procedural fairness perceptions ($\mathbf{H}_{3.1b}$; $F(1,254) = 0.004$, $p = 0.95$, $\eta_p^2 < 0.01$) and a Bayesian analysis returned moderate evidence in favor of the null hypothesis that human oversight has no effect here ($\mathrm{BF}_{01} = 7.43$). However, there was a strong effect of *contestability* ($\mathbf{H}_{3.1c}$; $F(2,254) = 20.60$, $p < 0.001$, $\eta_p^2 = 0.14$; $\mathrm{BF}_{10} > 1000$). We further found no evidence in favor of the effect of *contestability* on decision subjects' perceptions of procedural fairness being moderated by the presence of *explanations* ($\mathbf{H}_{3.1d}$; $F(2,254) = 0.16$, $p = 0.85$; $\eta_p^2 < 0.01$, $\mathrm{BF}_{01} = 12.95$) or by the presence of *human oversight* ($\mathbf{H}_{3.1e}$; $F(2,254) = 0.005$, $p = 1.00$; $\eta_p^2 < 0.01$, $\mathrm{BF}_{01} = 13.35$). We also did not find any evidence of an interaction between *task stakes* and *human oversight* ($\mathbf{H}_{3.2b}$; $F(1,254) = 0.06$, $p = 0.80$, $\eta_p^2 < 0.01$; $\mathrm{BF}_{01} = 7.32$) or *task stakes* and *contestability* ($\mathbf{H}_{3.2c}$; $F(2,254) = 0.52$, $p = 0.60$, $\eta_p^2 < 0.01$; $\mathrm{BF}_{01} = 7.20$) when predicting perceptions of procedural fairness.

We performed a multiple linear regression analysis to test the association of informational and procedural fairness perceptions with overall fairness perceptions ($R^2 = 0.46, F(3,263) = 76.02, p < 0.001$). Our results show that *perceptions of informational fairness* ($\mathbf{H}_{3.3a}$; $\beta = 0.27$, $p < 0.001$) and *perceptions of procedural fairness* ($\mathbf{H}_{3.3b}$; $\beta = 0.87$, $p < 0.001$) both predicted overall fairness perceptions, with procedural fairness perceptions being the stronger predictor. However, we did not find evidence that perceptions of informational and procedural fairness interact ($\mathbf{H}_{3.3c}$; $\beta = -0.09$, $p = 0.07$) when predicting overall fairness perceptions.

In sum, we found evidence in favor of four of our hypotheses: $\mathbf{H}_{3.1a}$, $\mathbf{H}_{3.1c}$, $\mathbf{H}_{3.3a}$, and $\mathbf{H}_{3.3b}$, indicating effects of explanations on informational fairness perceptions and contestability on procedural fairness perceptions, respectively (Figure 4.3). We also show that informational and procedural fairness perceptions are positively related to overall fairness perceptions.

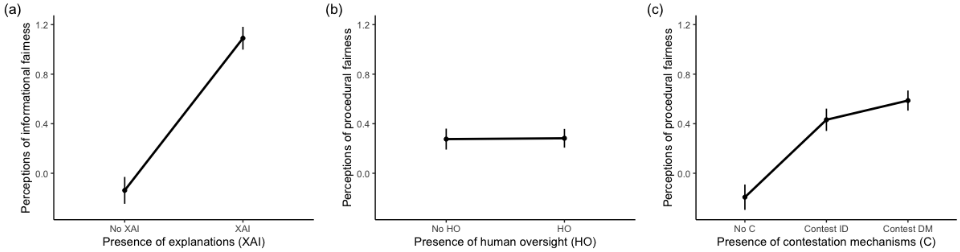


Figure 4.3: Effects of (a) explanations on perceptions of informational fairness and, (b) human oversight, and (c) contestability on perceptions of procedural fairness (HO = human oversight, C = contestability, ID = initial decision, DM = decision-maker). Connecting lines are used to visualize the differences between the cases.

4.5.3. Exploratory Analyses

In addition to the hypothesis tests (see Section 4.5.2), we performed several exploratory analyses to better understand our results and identify any unforeseen but interesting trends in our data. Note that these are not confirmatory results as we did not preregister any of the analyses presented in this subsection.

Decision tasks are subjective and personal [211], so we conducted a manipulation check regarding the stakes of the task. We performed a t-test between the pre-defined task stakes (low for a holiday loan, high for a home loan) and participants' perceived task stakes. Our results indicate that the holiday loan ($M = 0.38$, $SD = 1.31$) was, indeed, regarded as a lower-stakes scenario compared to the home loan ($M = 1.70$, $SD = 1.07$; $t(258.61) = 9.09$, $p < 0.001$).

Because contestability is composed of three different groups, we performed pairwise comparisons to analyze the specific differences with respect to procedural fairness perceptions. We observed no significant difference between the effect that the two suggested contestation mechanisms have on procedural fairness perceptions (Tukey-adjusted $p = 0.45$), but both of them differed from the option with no contestability (Tukey-adjusted $p < 0.001$ in both cases).

We also looked at the effects of explanations, human oversight, and contestability on the sub-elements of informational and procedural fairness perceptions. Each of these sub-elements is assessed by one individual item in the fairness perception questionnaires. For *informational fairness perceptions*, we evaluated whether participants thought that Kim received (1) thorough, (2) reasonable, (3) tailored, and (4) understandable information. For *procedural fairness perceptions* we evaluated perceptions of (1) procedural voice, (2) influence over the outcome, (3) consistency of the process, (4) lack of bias, (5) accuracy of factors, (6) correctability, and (7) ethicality. We thus performed multi-way ANOVAs with explanations, human oversight, contestability, and task stakes as between-subjects factors, and the sub-elements that compose informational and procedural fairness perceptions as the dependent variables.

Effects of Explanations.

As expected, providing explanations had a positive effect on decision subjects' perceptions of informational fairness. Participants considered that, whenever explanations were added, the bank was giving thorough ($F(1,249) = 104.00$, $p < 0.001$, $\eta_p^2 = 0.29$) and reasonable ($F(1,249) = 40.31, p < 0.001$, $\eta_p^2 = 0.14$) information that would make Kim understand ($F(1,249) = 19.84, p < 0.001$, $\eta_p^2 = 0.07$) the way in which the decision was made. Participants also considered that these explanations were tailored to Kim's needs ($F(1,249) = 45.55, p < 0.001$, $\eta_p^2 = 0.15$). The effect on procedural fairness was partial: our exploratory analysis suggests that explanations affected perceptions of process consistency ($F(1,254) = 16.80, p < 0.001$, $\eta_p^2 = 0.06$), potentially because explaining to decision subjects how each factor contributes to a final decision may make them discover that the process is standardized and uses the same criteria for every client. Explanations also seemed to interact with contestability in perceptions of procedural consistency ($F(2,254) = 3.83, p < 0.05$, $\eta_p^2 = 0.03$). Moreover, we checked the interaction of AI literacy and explanations on informational fairness perceptions by performing a multiway ANOVA with explanations, human oversight, contestability, task stakes, and AI literacy as between-subject factors and perceived informational fairness as the dependent variable. We found that *AI literacy* may have an effect on perceptions of informational fairness ($F(1,249) = 4.14, p < 0.05$, $\eta_p^2 = 0.02$) and that *explanations* and *AI literacy* may interact ($F(1,249) = 4.19, p < 0.05$, $\eta_p^2 = 0.02$) in creating perceptions of informational fairness (see Figure 4.5). These results suggest that participants with low AI literacy rated informational fairness perceptions negatively when no explanations were given, but their perceptions of informational fairness substantially increased when decisions were explained. The presence of explanations had a milder effect on informational fairness perceptions among participants with higher AI literacy.

Effects of Human Oversight.

Our exploratory analyses suggest that human oversight had no effect on any of the items that contribute to procedural fairness perceptions individually. As a matter of fact, our results show that the inclusion of human oversight in the initial decision has a slight negative impact on perceptions towards process consistency and lack of bias (Figure 4.4). Human oversight and contestability further seemed to interact in affecting procedural voice perceptions ($F(2,254) = 4.08, p < 0.05$, $\eta_p^2 = 0.03$) and outcome influence ($F(2,254) = 3.65, p < 0.05$, $\eta_p^2 = 0.03$). This result may suggest that configurations where decision subjects can contest the decision basis of the process lead to varying degrees of procedural voice and outcome influence perceptions depending on whether the initial decision was overseen by a human or not.

Effects of Contestability.

In our exploratory analysis, we found that contestability mainly contributed to the "correctability" sub-element of procedural fairness perceptions ($F(2,254) = 108.29, p < 0.001$, $\eta_p^2 = 0.46$). This is somewhat unsurprising considering that correctability directly refers to the requirement of having an appeal process in place [253]. Interestingly, however, although contestability seemed to improve perceptions of procedural voice ($F(2,254) = 13.76, p < 0.001$, $\eta_p^2 = 0.1$), the mean values of perceived procedural voice are still below zero (on a $[-3,3]$ scale) for all three configurations: the configuration where there is no contestability ($M = -1.84, SD = 0.16$), the configuration where participants can contest the initial decision ($M = -0.81, SD = 0.17$) and the configuration where participants

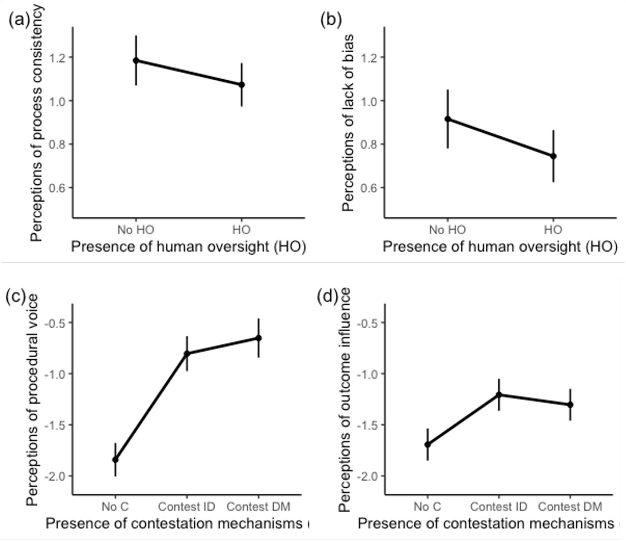


Figure 4.4: Effects of human oversight on perceptions of (a) process consistency and (b) lack of bias; effects of contestability on perceptions of (c) procedural voice and (d) outcome influence (HO = human oversight, C = contestation, ID = initial decision, DM = decision-maker). Connecting lines are used to visualize the differences between the cases.

can contest the decision-maker ($M = -0.65, SD = 0.19$) (Figure 4.4). The mean values for perceptions of outcome influence are also below zero for all three configurations: no contestability ($M = -1.69, SD = 0.16$), contest initial decision ($M = -1.21, SD = 0.16$) and contest decision-maker ($M = -1.30, SD = 0.16$). This suggests that none of the contestation mechanisms put in place may sufficiently contribute to decision subjects' sense of having a voice in the process and influence over the outcome (i.e., the first two sub-elements that constitute procedural fairness perceptions). Our exploratory results also do not point to any differences between contestation types for any of the sub-elements that compose procedural fairness perceptions; except for ethicality ($\beta = -0.81, p < 0.05$). This might indicate that, based on ethical and moral standards, participants do require human intervention in the review process. Note that there is no interaction between contestation types and human oversight for ethicality, which could suggest that having a human-in-the-loop configuration in the original decision is no substitute for human intervention in the review process when upholding ethical standards.

Effects of Task Stakes.

Our exploratory analyses surprisingly suggest that task stakes contribute to one item of procedural fairness perceptions: adequacy of factors (e.g., credit score, loan amount requested, total annual income) ($F(1,254) = 86.79, p < 0.001$, $\eta_{\text{p}}^2 = 0.25$; see Figure 4.5). This suggests that participants perceived the decision factors used in our scenario as less adequate for the low-stakes decision (holiday) than for the high-stakes decision (buying a house).

	Informational Fairness					Procedural Fairness								Overall Fairness
	Mean	Th	R	T	U	Mean	V	Inf	Cnst	LB	AF	Crr	Eth	
Explanations	***	◇	◇	◇	◇	◇		◇	◇				◇	
Explanations × Task Stakes														
Explanations × AI literacy	◇		◇											
AI literacy	◇				◇									
Human Oversight														
Human Oversight × Task Stakes														
Contestability						***	◇				◇	◇		
Contestability × Explanations									◇					
Contestability × Human Oversight		◇					◇	◇						
Contestability × Task Stakes											◇			
Task Stakes											◇			
Informational Fairness Perceptions														***
Procedural Fairness Perceptions														***

Table 4.3: Summary of our results. *** refer to confirmatory results ($p < 0.001$), whereas ◇ refer to exploratory results ($p < 0.05$). Empty cells indicate an absence of significant effect between variables. Mean = averaged value of the sub-items that constitute faceted fairness perceptions, Th = Thorough, R = Reliable, T = Tailored, U = Understandable, V = procedural Voice, Inf = Outcome Influence, Cnst = process Consistency, LB = Lack of Bias, AF = Adequacy of Factors, Crr = Correctability, Eth = Ethicality.

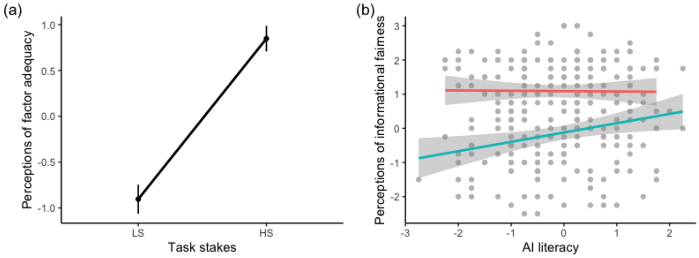


Figure 4.5: (a) Effect of task stakes on perceptions of factor adequacy (LS = Low stakes, HS = High stakes). (b) Interaction between explanations and self-reported AI literacy on perceptions of informational fairness. Red refers to the configurations where explanations were given and Green refers to the configurations with no explanations. Connecting lines are used to visualize the differences between the cases.

4.5.4. Qualitative Analysis

We performed our qualitative analysis using a reflexive thematic analysis [68]. We inductively generated individual codes from the responses our participants gave to the open-ended questions and we then clustered them into **code groups**. We identified three main tension areas: one related to perceptions of informational fairness and two related to perceptions of procedural fairness. This section explains each of those areas of tension in detail. For a comparison and discussion between quantitative and qualitative results, see Sections 4.6.1, 4.6.2, and 4.6.3. We again refer to quotes as Q.*i*, where *i* is the index of a specific quote. Appendix B.1. shows all selected quotes.

Tension #1: Amount of Information vs. Generating Understanding for All.
Our qualitative results indicate that getting detailed information about the decision was a general concern among participants. Participants who were placed in a configuration without explanation of the decision outcome directly highlighted the need for the bank to give **detailed explanations** (115/133) about the way in which different factors are used for making the decision and the reasons for the outcome (Q.11). They also considered that the bank should provide decision subjects with an alternative **course of action** (34/133; Q.12).

Participants who were placed in scenarios where the bank would offer explanations of the decision outcome positively evaluated the level of detail of this information (70/134). They generally also appreciated the fact that the counterfactual scenarios gave actionable information (21/134). Some of them requested **further information about the process and the algorithmic system** itself (51/134; Q.13). However, some participants pointed out that increasing the amount of information could generate **difficulties in understanding** (23/134) the explanations and could restrict such understanding to people with literacy in AI (Q.14).

Tension #2: Human Involvement vs. Timely Decision-Making.
Another major theme in our qualitative analysis was that of human involvement. Our qualitative analysis suggests that, regardless of the presence or absence of human oversight, participants were still asking for a **higher degree of human involvement** (75/267) in the process (e.g., by including a human that deals with borderline cases, or by allowing

decision subjects to personally interact with a bank employee). In cases where human oversight was included in the original decision, our participants thought that this would ensure reliability. However, some (13/267) of them indicated that a human should always make the final decision, for every instance (Q.15, Q.16).

On the other hand, as some of our participants highlighted, not having humans involved could make the **process speedy** (47/267) and would allow Kim to explore alternative options (Q.17). Although we did not explicitly compare the difference in time of having a human or an algorithmic system (with or without human oversight) making the decision, the presented scenario did mention that the reason for introducing algorithmic decision-making processes was due to time constraints. Many participants referred to the temporal dimension as one that makes the process fair (Q.18, Q.19).

Tension #3: Standardized Fact-based Process vs. Accounting for Personal Circumstances.

The fact that an algorithmic system was fully or mainly driving the process also encouraged reflections on the advantages and disadvantages of having a standardized process that treats **everyone equally** (44/267; Q.20). Some of our participants considered that introducing algorithmic systems in decision-making processes helps to **get rid of human biases** (39/267). They considered that thanks to such systems, the process would not be subject to human subjectiveness and prejudice (Q.21). Introducing an algorithmic system was also viewed as contributing to the consistency of the decision-making process. Participants generally appreciated that the same information was considered for everyone (Q.22). The basis of the decision-making process was also regarded as sound because it was **based on facts** (40/267; Q.23). Some (27/267) indicated that the bank should consider additional factors when making a decision, but, in general terms, the presented factors were considered fair and relevant (Q.24).

Despite the general sentiment of facts being a sound basis for decision-making, some of our participants highlighted the need to sometimes consider **individual circumstances** (17/267; Q.25, Q.26). Humans were viewed as being more flexible and prone to give in to cases that are close to the decision boundary (Q.27). Some participants pointed out that a human should be responsible for double-checking boundary cases (Q.28). In those cases, participants requested the implementation of negotiation mechanisms (Q.29) that would allow decision subjects to **discuss with humans** (47/267; Q.30) who could treat the situation with compassion (Q.31).

4.6. Discussion

In this section, we relate quantitative results with qualitative ones and reflect on our key findings. Each subsection summarizes the results related to one of the tested factors and its entanglements (i.e., explanations in Section 4.6.1, human oversight in Section 4.6.2, and contestability in Section 4.6.3). We also list the practical implications of our findings, highlight future challenges, and reflect on the benefits and shortcomings of adopting a multi-dimensional approach for capturing perceptions of fairness (Section 4.6.4). We finally acknowledge the limitations of our study (Section 4.6.5).

4.6.1. Leveraging Transparency Beyond Outcome Explanations

Our quantitative results show that explanations improve informational fairness perceptions (see Section 4.5.2). Exploratory findings further suggest that AI literacy may moderate the effect of explanations on informational fairness perceptions, i.e., indicating that the effect of explanations on informational fairness perceptions is stronger for participants with low AI literacy (see Section 4.5.3 and Figure 4.5). However, contrary to our expectations, and to suggestions from earlier work [365], we did not find evidence that explanations moderate the effect of contestability on procedural fairness, i.e., help participants question structural aspects of the decision-making process such as the factors requested by the bank and how these are used. The insights we obtained from our qualitative analysis suggest that participants were generally happy with the factual basis of the decision in question (see Section 4.5.4). It should be noted that, as opposed to earlier work [365] and our own preliminary study, we had decided to discard gender as one of the decision-making factors in our main study because it is explicitly protected by law [46]. This might have influenced how people perceived the decision basis. Moreover, some participants were asking for system-level explanations that would enable them to explore and evaluate biases encoded in the algorithmic system. The lack of this information might have prevented them from questioning additional aspects of the decision-making.

Implications. Although our study replicated the finding from earlier work that explanations support informational fairness perceptions [365] (which in turn contribute to overall fairness perceptions), restricting explanations to technical solutions that are currently available through XAI may limit the grounds for contestations [269]. Our results (e.g., Q.13) suggest that providing decision subjects with information that goes beyond outcome explanations could support contestations that are not only limited to post-decision mechanisms but that apply to the system lifecycle as a whole [8]. These system-level explanations could include information about data, algorithmic features, or the way in which algorithmic systems are integrated in broader workflows [116]. For instance, previous studies have shown that data-centric explanations [21] have the potential to assist decision subjects in assessing fairness. Future work should look into explanations and transparency that go beyond outcomes and test how these insights affect perceptions of informational fairness and whether they set grounds for contestations that go beyond appeal processes. We foresee that this would not only have implications for perceptions of informational fairness but also for perceptions of procedural fairness.

Challenges. Previous research has demonstrated that increasing levels of transparency can lead to information overload [84], so expanding explanations could restrict understanding to individuals with literacy in AI. Moreover, earlier work has pointed to a risk that malicious actors might use explanations to defraud algorithmic systems [421] or to manipulate decision subjects by conveying untruthful levels of "fairness" [271]. Future work should look into methods for designing strategies that leverage adequate levels of transparency [421] and that convey *appropriate fairness perceptions* (i.e., condition that is satisfied if fairness perceptions towards a system are high when the system is indeed fair) [364]. Such strategies should be adapted to decision subjects' insight needs [386] and designed in a way that they would understand [29, 222]. For example, these could include videos [406], stories [406], or comics [418, 420]. Our qualitative analysis further revealed some participants' feeling that the process could not be biased because it is impossible for algorithmic systems to be biased (Q.20), suggesting that future explanations should also account for decision subjects' imaginaries [297] and expectations [228] around algorithmic systems.

4.6.2. Designing Appropriate Human-AI Configurations

Our quantitative results do not contain any evidence that human oversight would affect decision subjects' procedural fairness perceptions; in fact, a Bayesian analysis even revealed moderate evidence that human oversight has no effect here (see Section 4.5). These results resonate with earlier work on the topic [419], where a case-by-case human intervention did not contribute to perceptions of fairness. Nevertheless, our qualitative results suggest that, regardless of human oversight in the original decision, participants were still asking for a higher degree of human intervention (e.g., Q.15; see Section 4.5.4). The reason for this might be that end users might think about the decision-maker in binary terms, as either "a human" or "not a human" [239]. Since, even in the scenario with human oversight, the first prediction was made by the algorithmic system, our participants might still have thought about it as a non-human decision-maker. This would explain why human oversight did not affect perceptions of procedural fairness and why, even in the case where the decision was overseen by a human, participants were asking for more human intervention in the process.

Implications. More research is needed to find adequate forms of human-AI collaborations in algorithmic decision-making processes. Future studies should go beyond configurations where humans confirm the quality of the decision made by an algorithmic system [14] and craft alternative human-AI teams. For instance, AI systems could access large quantities of data and perform preliminary analyses to produce easily digestible summaries for human experts to make final decisions [327]. Such a configuration would respond to our participants' desire to always have a human making the last decision. A follow-up study to ours could test perceptions towards human decision-making processes that are advised by AI systems [39, 440] rather than algorithmic decision-making processes that are overseen by humans. One could argue that many studies have already studied different human-AI teaming configurations. However, these studies have mainly focused on exploring the interaction of data domain experts (i.e., bank employees in our case) with algorithmic systems and distilling the effect on trust [318, 361] or trust-related constructs [413] such as reliance [331, 440]. Future studies should also capture decision subjects' fairness perceptions for each of those configurations.

Challenges. Including humans in algorithmic decision-making processes costs time [80, 114, 271] and our qualitative results suggest that participants value timely decision-making processes. For appeal processes, Lyons et al. [271] found that, when subject to a trade-off situation, participants prioritised the type of review and the review time rather than the reviewer. We emphasize the need to perform more studies where participants are shown the time cost of different configurations so as to capture their perceptions of procedural fairness in a space of trade-offs. Furthermore, our participants regarded configurations with no human intervention as less biased and more consistent. We echo Almada [14] and suggest that comparative measures of performance of human-controlled and fully-automated procedures should be included. This would allow end users to freely shape their preferences and fairness perceptions in an informed way.

4.6.3. Giving Voice to Decision Subjects

As we hypothesized, our quantitative results show that including contestability (in the form of appeal processes) enhances decision subjects' perceptions of procedural fairness. Our qualitative results back up the value that participants put on the ability to contest the decision. Despite the positive effect of contestability on perceptions of procedural fairness, perceptions of procedural voice and influence over the outcome were still negative. In a within-subjects user study, Lyons et al. [271] found that participants perceived the *new information* appeal condition (equivalent to our "option to contest the initial decision and provide additional information" appeal condition) as fairer than the rest of the suggested appeal processes. Contrary to these findings, we do not find any differences between the suggested appeal processes. This might be due to the between-subject nature of our study. Lyons et al. [271] also found that the reason for the preference towards this condition was that decision subjects perceived they had a "voice" in the decision-making process. Our results contradict these findings, and indicate that, even when any of the suggested appeal processes are in place, our participants did not have the feeling that the decision subject had a voice in the process or influence over the outcome. This discrepancy might be due to the nature of the performed analysis. Lyons et al. [271] arrived at this conclusion through a thematic analysis of qualitative data, whereas our results rely on quantitatively evaluating responses to statements that directly address perceptions of procedural voice and influence over the outcome.

Implications. Our findings highlight that, although contestability enhances decision subjects' perceptions of procedural fairness (which in turn contribute to overall fairness perceptions), more research in contestable AI is needed. The field of contestable AI is still growing [9] and many of the guidelines on how to design for contestability are conceptual in nature [9, 179, 269]. Further research is necessary to translate those conceptualizations into actual design guidelines [8, 249] and validate designs of contestable algorithmic systems. Our results also suggest the need to research into the design of contestation mechanisms that effectively provide voice and outcome influence to decision subjects. Sarra [355] argue that a "dialectical exchange" is necessary between decision subjects and human reviewers to effectively support contestability. This resonates with our qualitative findings: many of our participants were asking for options to personally discuss or negotiate the outcomes with humans. Our participants considered that discussing the decision with humans would potentially lead to a change in outcome for cases that were close to the decision boundary (e.g., Q.27, Q.28; in line with earlier work [148, 271]) and that humans would treat decision subjects with dignity and compassion (e.g., Q.31; also in line with previous research [59, 271, 407]). These findings further suggest that contestations might be better designed as dialogues [179, 224], rather than mere appeal processes. When it comes to outcome influence, future research should focus on ways of increasing the ability of subjects to exercise agency and true influence over the process [23]. This entails allowing decision subjects to determine the input data that they want to provide along with the ability to influence the logics of the decision-making process [249]. A promising research line in this field is that of *interac-*

tive contestations [185].

Challenges. A major challenge when trying to give effective outcome influence to decision subjects is the distribution of levels of control across individuals. Since the process will eventually influence multiple people rather than one individual, the way in which this control is distributed remains a key challenge [304]. We consider that participatory design strategies [176], such as the workshops conducted by Vaccaro et al. [407], can help deal with the trade-offs identified in our qualitative analysis. These workshops facilitate conversations among different stakeholders (e.g., the development team and decision subjects) and could, therefore, help identify the compromises in designing contestation mechanisms that attend to individual circumstances while contributing to perceptions of process consistency.

4.6.4. Multi-dimensional Measurement of Fairness Perceptions

In this chapter, we advocated for a multi-dimensional approach for capturing perceptions of fairness, inspired by literature in human decision-making. Our quantitative analyses confirm that informational and procedural facets of fairness predict overall fairness perceptions. Moreover, this multi-dimensional approach has enabled us to perform exploratory analyses that have generated a nuanced understanding of how people perceive each algorithmic configuration. Our findings, therefore, suggest that future studies and practical applications could benefit from adopting a multi-dimensional rather than a one-dimensional approach.

Despite our promising findings, using a tool that was designed for human decision-making to evaluate algorithmic decision-making may not encompass the unique challenges that the inclusion of algorithmic systems bring to existing processes (as it is the case for other fields such as human-agent collaboration [88]). Our aim behind using this tool designed for human decision-making in an algorithmic context was to distill insights from it and to identify future research directions. There is evidence that suggests that decision subjects care about justice-related aspects in algorithmic decision-making, as they care in human decision-making [59]. However, we acknowledge that there are novel considerations that the usage of these systems results in [59] and that future work should consider. For instance, the approach suggested by Colquitt [95] does not explicitly include the temporal dimension of the decision-making process as an attribute that contributes to perceptions of procedural fairness. Through our qualitative analysis, we found that this aspect was paramount for our participants. We note that most of the criteria we evaluated were defined several decades ago. Due to societal changes and a change in perceptions of time brought in by algorithmic systems, further research would be needed to consider and effectively evaluate speed of decision-making as a procedural justice principle [409]. We, therefore, encourage further research into defining standardized methodological approaches that appropriately capture perceptions of fairness across dimensions while being specifically adapted to algorithmic decision-making.

4.6.5. Limitations

In this section, we summarize limitations of our study that could represent threats to its validity.

Reflections on our experimental setting. The design of our study might have had an impact on the obtained results. First, the between-subjects nature of the study might have prevented participants from comparing different algorithmic configurations. The effects of task stakes and human oversight might have been diluted because of this. Second, the scenario used for conducting our controlled user study presented a case that participants considered to be close to the decision boundary (see Q.27). This made the request to have a human involved in the decision-making process, for example, to be especially relevant for some participants (see Q.28). Fairness perceptions and the desires expressed by participants might have been different if we had included scenarios with different characteristics. Third, the design of our experiment described a loan denial scenario for an individual called Kim. As opposed to some other authors (e.g., [271, 419]) we decided to tell this story in the third person [23, 365, 381] with no reference to the individuals' personal characteristics. The reason behind this design choice was to minimize, as far as possible, the *outcome favourability bias* [419]. In the same line, we limited the interaction between participants and the algorithmic system to a one-shot interaction. Previous research has shown that, under repeated interactions, system favorability towards the group that the decision subject belongs to has an effect on fairness perceptions [154]. Our results indicate that, generally speaking, participants were happy to endorse negative outcomes if explanations and contestation mechanisms were in place. However, outcome favourability bias might have resulted in different reactions had we referred to a case where the participants themselves had been denied a loan or had we disclosed the demographics of different individuals and asked participants to repeatedly interact with the algorithmic system. Fourth, although we varied the level of stakes involved in the task and found that perceptions of informational and procedural fairness are robust across stakes, our study is still limited to a loan decision-making scenario. Results may vary depending on the context. Fifth, terminology has been claimed to affect decision subjects' fairness perceptions [239]. Langer et al. [239] suggest that the usage of multi-item measurement tools softens the impact of terminology, an advice we followed when measuring perceptions of informational and procedural fairness. However, results may have been different had we used terms such as *algorithmic system*, *statistical model*, or *computing system* instead of *artificial intelligence*.

Generalizability across cultures. For our study we recruited participants from the Global North whose first language was English. Previous work has shown that cultural and geographical differences play a key role in perceptions towards algorithmic systems [28, 54, 211]. Thus, we acknowledge that our study is subject to representativeness limitations [252].

Need to incorporate empirical ethics as part of broader design frameworks for algorithmic systems. Empirical studies represent a necessary strategy for testing the practical implications of theoretical claims. However, moving towards algorithmic decision-making processes that enhance decision subjects' feelings of justice requires that empirical studies are part of broader efforts to create methodological tools that consider different stakeholders' (including decision subjects) viewpoints in the design and evaluation processes of algorithmic systems [334, 438].

4.7. Chapter Takeaways

In this chapter, we presented a preregistered user study investigating the *right to contest* automated decisions and related safeguards as interpreted from Article 22(3) of the General Data Protection Regulation (GDPR) [131]. We evaluated the effect that the presence (or lack thereof) of explanations, human oversight, and appeal mechanisms on decision subjects' ' informational, procedural, and overall fairness perceptions. We found that explanations and appeal mechanisms affect perceptions of informational and procedural fairness, respectively. Even if appeal mechanisms positively contributed to procedural fairness perceptions, they were still negatively perceived. We did not find evidence of the effect of human oversight on these measurements. We also found that perceptions of informational and procedural fairness, independently, are positively related to perceptions of overall fairness. Our results confirm the need to rethink traditional appeal mechanisms when dealing with algorithmic decision-making. The findings presented in this chapter additionally highlight the need to further look into the operationalization of human intervention and the effect that this safeguard has on decision subjects' fairness perceptions (**RQ4**).

5

Decision Subjects' Fairness Perceptions Towards Human Intervention

In this chapter, we evaluate how decision-making configurations with varying levels of human intervention affect decision subjects' fairness perceptions in algorithmic decision-making (**RQ4**). Studying the effect of human intervention on decision subjects' fairness perceptions is relevant to contestability because human intervention is one of the safeguards that conditions decision subjects' right to contest automated decision as interpreted from Article 22(3) of the General Data Protection Regulation (GDPR). While human intervention has been claimed to protect decision subjects' rights, it is not clear how decision-subjects perceive hybrid decision-maker configurations (i.e., combining humans and algorithms). We address this gap through a mixed-methods 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*, *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 on 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.

This chapter is based on the following publication: **Mireia Yurrita**, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia Pont, Alessandro Bozzon. "Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions". In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25)

5.1. 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 [354]. Human intervention is included in regulatory efforts, like the European Union's General Data Protection Regulation (GDPR) [131], 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 [140, 230, 339]. 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) [59, 271, 437]. Crafting algorithmic decision-making processes that uphold decision-subjects' standards of fairness is, in turn, key to ensuring the responsible implementation and broader acceptance [53, 119, 247, 316] of AI systems that could help deal with large-scale, increasingly-complex issues [140].

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., [23, 80, 181, 247, 264, 271, 317, 406]). In most cases, these studies concluded that decision-subjects prefer exclusively-human configurations [80, 247, 264]. 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 [419, 437]. 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., [21, 23, 247, 365]) 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

[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*.

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 chapter, 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 5.1). To this end, we adopted a mixed-methods approach [197] grounded in a context of algorithmic policy 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 5.3) — decision-subjects of illegal holiday rental detection. Note that these interviews are the same ones we analyzed in chapter 3. However, by re-analyzing these interviews with a focus on decision-makers rather than contestability, we 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) [280]. We chose to characterize perceptions towards decision-makers through the ABI model [280] because this model distinguishes perceptions of trustworthiness towards decision-makers from trust (see section 5.2.3). The following research questions guided the interview study:

   - **RQ4.1.1.:** What are the main characteristics that decision-subjects consider when assessing the adequacy of decision-maker configurations?
   - **RQ4.1.2.:** 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 (**RQ4.1.1.**). We mapped these characteristics onto the Ability, Benevolence, and Integrity (ABI) model [280] (**RQ4.1.2.**).

[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)

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 5.5). 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:

   - **RQ4.2.1.**: 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?
   - **RQ4.2.2.**: 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 (**RQ4.2.1.**). Our results also show that perceived ability and integrity positively relate to fairness perceptions (**RQ4.2.2.**). 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 chapter we, therefore, make two main contributions.

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

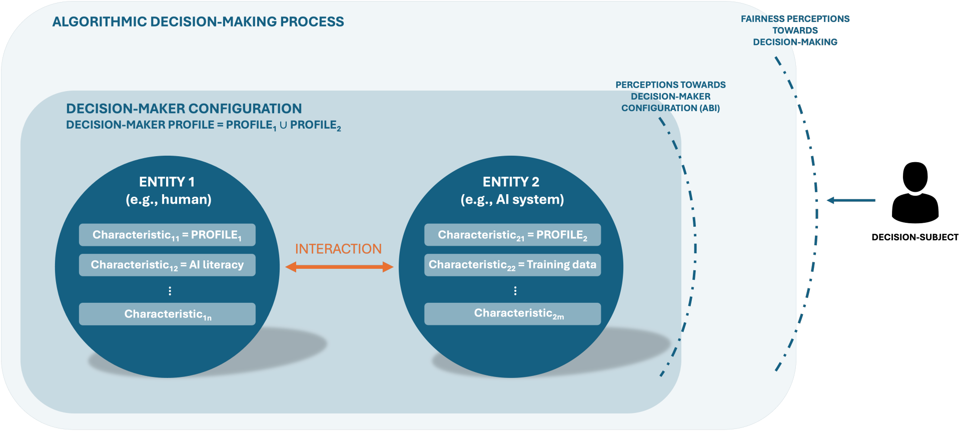

ALGORITHMIC DECISION-MAKING PROCESS
FAIRNESS PERCEPTIONS TOWARDS DECISION-MAKING
PERCEPTIONS TOWARDS DECISION-MAKER CONFIGURATION (ABI)
DECISION-MAKER CONFIGURATION
DECISION-MAKER PROFILE = PROFILE$_1$ ∪ PROFILE$_2$
ENTITY 1
(e.g., human)
Characteristic$_{11}$ = PROFILE$_1$
Characteristic$_{12}$ = AI literacy
⋮
Characteristic$_{1n}$
INTERACTION
ENTITY 2
(e.g., AI system)
Characteristic$_{21}$ = PROFILE$_2$
Characteristic$_{22}$ = Training data
⋮
Characteristic$_{2m}$
DECISION-SUBJECT

Figure 5.1: 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").

- 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.

All supplementary materials linked to this chapter can be found in our repository (https://doi.org/10.4121/8c19bb03-14de-4c85-b781-33eed0cac44a). 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.

5.2. Related work

This section first introduces the concept of *human intervention* for algorithmic decision-making (section 5.2.1). We then summarize recent research looking into fairness perceptions towards human intervention in algorithmic decision-making (section 5.2.2). 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 5.2.3).

5.2.1. 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) [131], as the act of providing human input by an individual with the competence and authority to change an algorithmic output [354]. In Article 22(3) of the GDPR [131], 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 [14]; a measure to acknowledge the "foundational indeterminacy of human self" [182, 286]. By allowing a competent human to provide input, human intervention is also claimed to be "an antidote to machine error" [14].

The role of human intervention is especially important in the public sector. Public decision-making processes deal with societally-sensitive topics [356, 445], where decision-subjects do not have an alternative to dealing with public administration [11] —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) [13, 445]. At the decision-making time, street-level bureaucrats engage in *reflexivity* [13] 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 [13]. Human intervention in the public sector aims at retaining and restoring street-level bureaucrats' discretionary power as part of algorithmic decision-making [445]. Given the relevance of human intervention in algorithmic decision-making in the public sector, we ground our study in a policy enforcement context.

5.2.2. 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 [55, 382].[5]

A considerable amount of work has been devoted to comparing fairness perceptions towards human *vs.* fully-automated decision-makers [23, 80, 114, 181, 233, 247, 262, 307, 317]. Most prior work has claimed that people normally considered humans to be more fair [80, 114, 181, 233, 247, 268, 307, 317]. Preference towards humans has been

[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., [38, 440]).

claimed to be caused by the perceived facility to convince them towards a favorable outcome as compared to algorithmic systems [148] and humans' ability to account for non-quantifiable aspects of the decision-making [307]. 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 [303, 419, 437] 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. [419] and Yurrita et al. [437] 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 [419] or those cases where the confidence of the AI output was low [437] (i.e., the interaction between the human and AI was based on *supervisory control* [348]). Nagtegaal [303], 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* [348], 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 [303]. Motivated by the absence of conclusive evidence, this chapter 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 [59, 116, 365, 437], the decision basis [166], details about the design [419] and data to train the system [21]. Explanations have been found to have a positive effect on informational fairness perceptions [365, 437], which, in turn, positively relate to overall fairness perceptions [437]. Using features that are perceived as relevant as the decision basis has been found to lead to positive fairness perceptions [166]. No evidence has been found of development procedures (e.g., developed in-house *vs.* outsourced) affecting fairness perceptions [419]. Information about the data used to train the system has been found to help users assess the fairness of a system [21].

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.

5.2.3. 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. [382] showed that fourteen of the reviewed studies directly captured fairness perceptions through single items. Instead, seventeen of the reviewed studies used fairness scales 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 [95]. Colquitt [95] 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's [95] 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 [280].

In organizational psychology, methods for capturing perceptions towards decision-maker configurations have instead been characterized as models of trust (e.g., [97, 280, 312]). Some scholars [97, 254, 281, 312, 341, 344] 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 [96]. An alternative line of work has studied trust as the trustor's willingness to be vulnerable to the trustee's actions [279, 280]. Mayer et al.'s [280] work is especially influential in this research area. Mayer et al. [280] 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 [280]. *Benevolence* refers to the goodwill of the trustee towards the trustor [280]. *Integrity* is defined as the trustor's perception that the trustee adheres to an acceptable set of principles [280]. 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.'s [280] model and adapted it to scenarios where the trustee is an automated agent [48, 231, 282].

In this chapter, 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 [96] and adopt the ABI model [280] to characterize perceptions towards decision-makers. The reason for adopting the ABI model [280] and not other trust models [97, 254, 281, 312, 341, 344] is that the ABI model [280] distinguishes perceptions of trustworthiness towards decision-makers from trust. The ABI model [280] 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 [96]. We apply the ABI model [280] in its original form. While studies in automation have shed light on how to adapt the ABI model [280] 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 [48], AI-enabled technology [168], AI for decision aid [380]). 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. [186]'s work has been the only one that characterized decision-makers from a decision-subject perspective for *algorithmic* decision-making. Höddinghaus et al. [186] characterized decision-makers through the original ABI model [280] 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.'s [186]'s approach and apply the original ABI model [186] with adapted ability items to capture perceptions towards decision-maker configurations. Unlike Höddinghaus et al. [186], 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., [398, 401]) have noted the existence of relationships between perceived trustworthiness and fairness perceptions. In an empirical study, Colquitt and Rodell [96] showed that the relationship between perceived trustworthiness towards decision-makers — conceptualized through the ABI model [280]— 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 [437]– 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 [280] dimensions.

5.3. Study 1: Qualitative Interview Study

In this chapter, we adopt a *mixed-methods approach* [197]. We followed prior work [44], 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 [44], 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 5.4 and 5.5).

5.3.1. Use case and Participant Recruitment

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) [45, 308]. 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.

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 [90], 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. We recruited participants by announcing our study in our institution and in short-term rental channels, and by reaching personal contacts.

Interview Procedure

In line with previous research (e.g., [70, 129, 211]), we used a scenario-based approach to introduce our participants to the use case.We introduced a fictional piece of news de-

[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)

scribing the use case 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) —see the interview protocol in Appendix B.1. 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.

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* [90, 91] 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.'s [280] 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 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.'s [280] definition of each dimension (e.g., we interpret empathy as "a positive orientation of the trustee towards the trustor" [280] 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 [90, 91] –see Figure 5.2. Unlike [44], in this chapter 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.

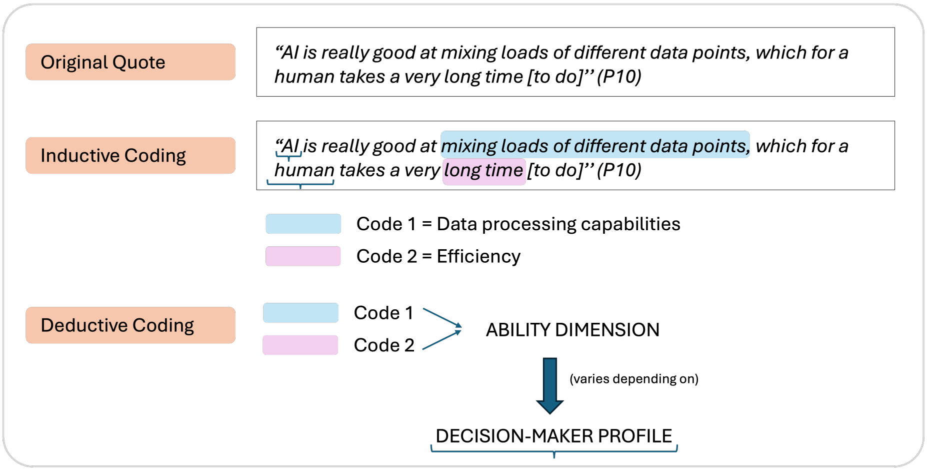

Original Quote
"AI is really good at mixing loads of different data points, which for a human takes a very long time [to do]" (P10)
Inductive Coding
"AI is really good at mixing loads of different data points, which for a human takes a very long time [to do]" (P10)
Code 1 = Data processing capabilities
Code 2 = Efficiency
Deductive Coding
Code 1
Code 2
ABILITY DIMENSION
(varies depending on)
DECISION-MAKER PROFILE

Figure 5.2: Example of the analysis process for one quote.

5.3.2. 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.'s [280] model (i.e., ability, benevolence, integrity).

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.

Perceived Benevolence

13 out of 21 participants evaluated the decision-maker's willingness to do good (i.e., *benevolence* [280]) 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).

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).

5.4. Hypotheses for Study 2

Combining the insights we got from our qualitative study with prior literature in algorithmic and human decision-making (e.g., [59, 96, 365, 419]), 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 5.3. All seven hypotheses were pre-registered before collecting the data. The combined effects between characteristics were examined in an exploratory fashion (see section 5.6.3).

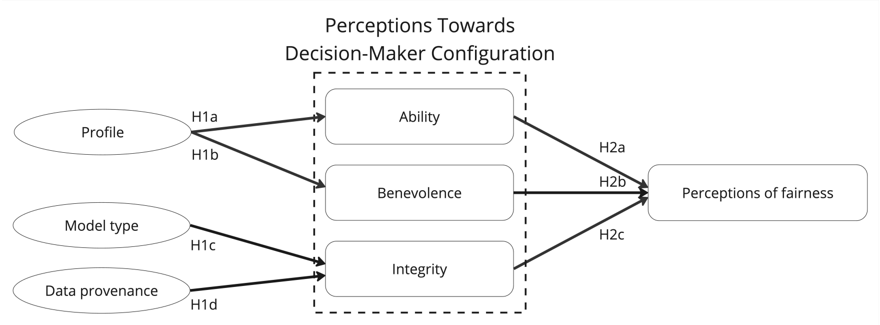


Figure 5.3: Overview of our hypotheses.

5.4.1. Hypotheses related to RQ4.2.1: Characteristics affecting Perceived Ability, Benevolence, Integrity

- **Hypothesis 4.1a ($H_{4.1a}$).** A hybrid decision-maker configuration (i.e., with human intervention)[9] 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 [247, 437]. However, these are also perceived to be less adaptable than humans [186]. 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 4.1b ($H_{4.1b}$).** A hybrid decision-maker configuration is perceived as more benevolent than a fully-automated one.

[9] 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 chapter, we will use the term "hybrid decision-maker configuration" *vs.* "fully-automated decision-maker configuration".

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, suggests that fully-automated decision-maker configurations are considered impersonal and dehumanizing [59]. 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 [264, 437]. In our qualitative study, participants highlighted that a decision-making process 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 4.1c ($H_{4.1c}$).** 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. [59], 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 et al. [419] 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 4.1d ($H_{4.1d}$).** 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 [21]. Even if the effect found by Anik and Bunt [21] 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.

5.4.2. Hypotheses related to RQ4.2.2: Effect of Perceived Ability, Benevolence, Integrity on Fairness Perceptions

- **Hypothesis 4.2a ($H_{4.2a}$).** 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 [96]. As opposed to these findings, we hypothesize that a difference in context might play a role. Colquitt and Rodell [96] 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 4.2b ($H_{4.2b}$).** Perceived benevolence relates positively to perceptions of fairness.

  Rationale. Prior literature in human decision-making found that for *benevolence* and *integrity*, the relationships between perceived trustworthiness and fairness perceptions are reciprocal; both influencing one another [96]. Similarly, we hypothesize that in *algorithmic* decision-making, *benevolence* will relate positively to fairness perceptions.

- **Hypothesis 4.2c ($H_{4.2c}$).** 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 [96]. We hypothesize that for *algorithmic* decision-making processes, there will also be a positive relation between perceived integrity and fairness perceptions.

5.5. Study 2: Large-Scale Quantitative Study

In this section, we describe how the insights generated in Study 1 (section 5.3) informed the design of our quantitative study. Our quantitative study aims at testing the hypotheses (see section 5.4) formulated based on the understanding we gained through the interview study.

5.5.1. Variables

Independent Variables

To capture perceptions towards decision-makers while avoiding *outcome favorability bias* [268, 419], 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

Table 5.1: 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 [419].
Model type	Probabilistic	Original system designed by Amsterdam municipality. Unlike prior work [419], 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 [118]).
	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)).

prior work [23, 365, 381, 437]. We generated $2 \times 2 \times 2 = 8$ different scenarios based on three independent variables.

- *Profile* (categorical, between-subjects). Each participant was randomly assigned to one of two configurations (Table 5.1):

  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 [303], decide whether to send a first warning to the property owner.[10]
  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 fol-

[10] 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 [40], the layout, wording and potential biases that we might trigger [118].

lowed by a different number of (+) signs to indicate that some of those parameters had a more prominent impact on the final probability [59, 116].

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.

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

- *Perceived ability*[11] (*continuous*). Measured by the average score on the six items suggested by Höddinghaus et al. [186].

- *Perceived benevolence* (*continuous*). Measured by the average score on the five items suggested by Mayer and Davis [279].

- *Perceived integrity* (*continuous*). Measured by the average score on the six items suggested by Mayer and Davis [279].

- *Perceived fairness* (*continuous*). Measured by a one-item construct on a 7-point Likert scale, following previous work [239, 247, 437].

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.

[11] 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. [186] and Mayer and Davis [279], 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 5.2: 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 [90], had an impact on perceptions towards decision-makers.
AI literacy	It has been shown to impact fairness perceptions in algorithmic decision-making [365, 437].
Affinity for technology	It has been shown to affect perceptions of ability towards algorithmic systems [239].
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 [233].
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 [247, 303].

- *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 5.2.

- *AI literacy* (*continuous*). Knowledge and expertise working or interacting with AI [365]. We captured it through the average score on the four items suggested by Schoeffer et al. [365].

- *Affinity for technology* (*continuous*). Curiosity towards and willingness to engage with the technical working of systems [239]. We captured it through the average score on the four items suggested by Franke et al. [143], following previous work [239, 437].

- *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 scale used by Kramer et al. [233] 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 used by Kramer et al. [233] 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 [288].

- *Perceived task complexity* (*continuous*). We measured perceived task complexity through a one-item construct on a 7-point Likert scale, similar to previous work [271, 437].

5.5.2. Procedure

We designed a four-step study —see Figure 5.4.

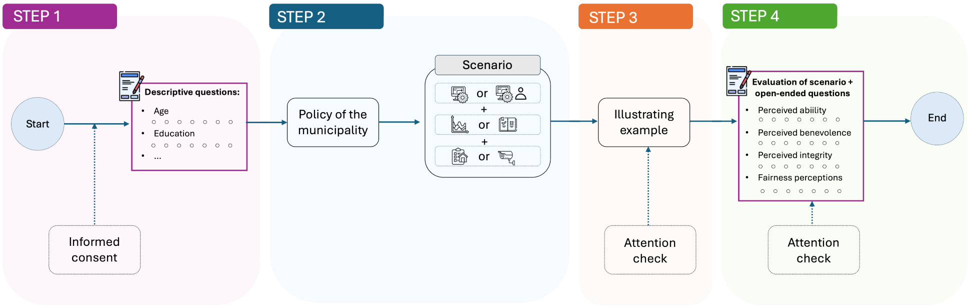


Figure 5.4: Procedure of the study.

Step 1. Participants accepted the informed consent and responded to questions related to our exploratory variables (see section 5.5.1).

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[12] 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 5.5.1). After each set of questions, participants were asked to further elaborate and justify their perceptions of ability, benevolence, and integrity through open-ended questions.[13] The second attention check was located between the questionnaire about perceived ability

[12] 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.

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

and perceived benevolence. Participants were finally asked to evaluate their fairness perceptions towards the algorithmic decision-making process.

5.5.3. 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* [136], 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 [211], 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.

5.5.4. 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 [172]. 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 [273] 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.

5.6. Results and Analysis of Study 2

In this section we summarize the quantitative —confirmatory (section 5.6.2) and exploratory (section 5.6.3)— 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.

5.6.1. Descriptive Statistics

For our study, we recruited 232 participants, out of which 223 participants passed both attention checks. Demographics are summarized in Table 5.3.

Table 5.3: 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%)

5.6.2. Hypothesis Tests

For our confirmatory analyses, we report the results for **H**$_{4.1a}$, **H**$_{4.1b}$, **H**$_{4.1c}$, **H**$_{4.1d}$ based on Kruskal-Wallis tests. To test **H4.2** we performed a non-parametric multiple linear regression.

H$_{4.1a}$: 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 5.5).

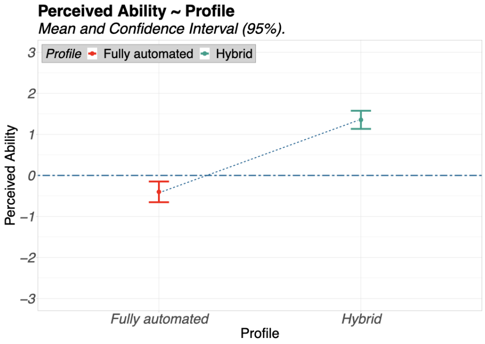


Figure 5.5: Effect of decision-maker's profile on perceived ability.

H$_{4.1b}$: 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 5.6). 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).

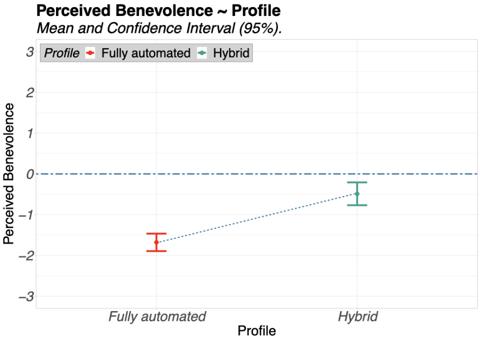


Figure 5.6: Effect of decision-maker's profile on perceived benevolence.

H$_{4.1c}$: We found no significant difference in perceived integrity across model type, $\chi^2(1) = 0.06$, $p > .1$, $\eta^2 = -0.004$.

H$_{4.1d}$: We found no significant difference in perceived integrity based on the input data provenance, $\chi^2(1) = 2.69$, $p = .1$, $\eta^2 = 0.008$.

H$_{4.2a}$, **H**$_{4.2b}$, **H**$_{4.2c}$: 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 (**H**$_{4.1a}$, **H**$_{4.1b}$, **H**$_{4.2a}$, **H**$_{4.2c}$). 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.

5.6.3. Exploratory Analyses

Besides the pre-registered confirmatory analyses, we also conducted exploratory analyses to better understand the observed effects. In particular, we performed two types of analyses: (1) additional main and interaction effects of the independent and control variables (see section 5.5.1) on *perceived ability*, *benevolence*, and *integrity*, and (2) mediation analyses as described earlier in section 5.5.4.

(1) Main and interaction effects.

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 5.7).

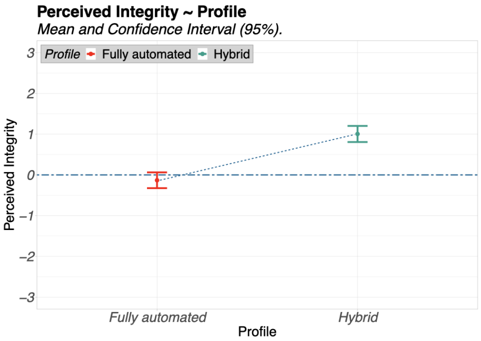


Figure 5.7: Effect of decision-maker's profile on perceived integrity.

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 5.8).

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.

(2) Mediation effects. We followed the procedure outlined by MacKinnon [273] 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 2.5th and

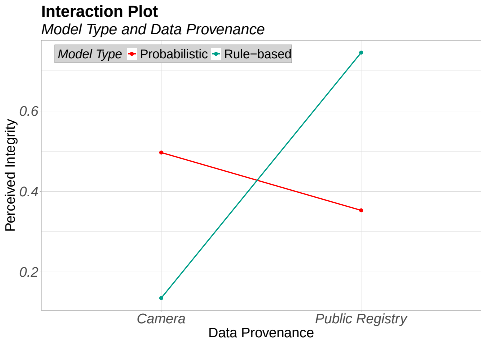


Figure 5.8: This figure shows the significant interaction effect between *model type* and *data provenance* when modeling *perceived integrity*.

97.5$^{\text{th}}$ percentiles.

Mediation Effect of Perceived Ability on the Relationship Between Decision-Maker's Profile and Perceived Fairness. In section 5.6.2 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 5.9) . In addition, we observed a complete and significant mediation effect, $\beta = 1.40$, $CI = [1.07, 1.77]$, $p < .001$.

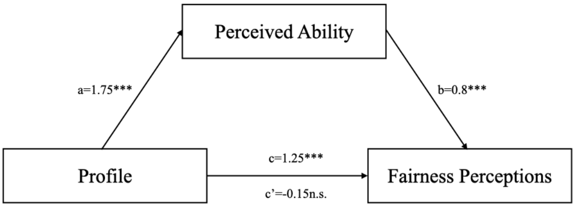


Figure 5.9: Mediation effect of *perceived ability* on the relationship between decision-maker's profile and fairness perceptions.

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 (β = 1.08, p < .001) were significant (see Figure 5.10).

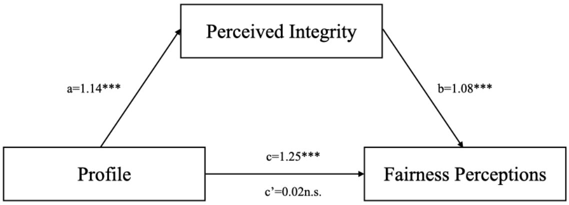


Figure 5.10: Mediation effect of *perceived integrity* on the relationship between the decision-maker's profile and fairness perceptions.

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, β = 0.45, p = .02. Therefore, we conducted a mediation analysis with perceived integrity as the mediator. Our results show a significant mediation effect, β = 0.20, CI = [0.05, 0.37], p = .008. The regression coefficients between policy agreement and fairness perceptions (β = 0.18, p = .02) and between perceived integrity and fairness perceptions (β = 1.09, p < .001) were significant (see Figure 5.11).

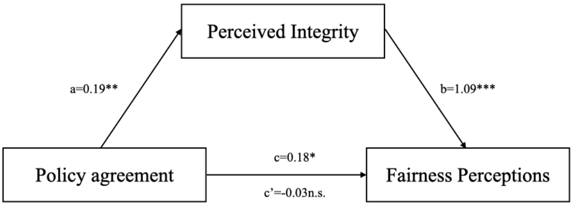


Figure 5.11: This figure shows the mediation effect of *perceived integrity* on the relationship between policy agreement and fairness perceptions.

5.7. 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.

5.7.1. Summary of Results In Relation to Previous Work

In this section, we summarize the results of our interview **(RQ4.1.1., RQ4.1.2.)** and large-scale quantitative studies **(RQ4.2.1., RQ4.2.2.)**. We focus on the findings related to the decision-maker profile in section 5.7.1, and on the findings related to the model type and data provenance in section 5.7.1.

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 [286]. 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 (**RQ4.1.1.**). Decision subjects evaluated the ability and benevolence of decision-maker configurations based on the decision-maker's profile (**RQ4.1.2.**). In our quantitative *between-subjects* study, and unlike previous work [303, 419, 437], 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 (**RQ4.2.1.**). 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 [348] and (2) differences in research method. Previous work comparing decision-subjects' perceptions towards fully-automated *vs.* hybrid decision-maker configurations [303, 419, 437] mainly gave humans a supervisory role and attributed them the task to monitor AI's actions (supervisory control [348]). 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 [348]). 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 [96]. Instead of capturing the effect of decision-maker configurations (a) on fairness perceptions directly [382], or (b) through fairness scales with little emphasis on the decision-maker [95], we first measured decision-subjects' 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 [11]. Unlike the private sector, where decision-subjects can, e.g., look for an alternative financial company if their loan gets rejected [437], decision-subjects necessarily have to deal with decisions made by public institutions [11]. 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.

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 (**RQ4.1.1.**) Participants evaluated decision-makers' integrity based on the model type and the input data provenance (**RQ4.1.2.**). 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 [179]). 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. [11]. Aljuneidi et al. [11] 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 [365] —without necessarily having to point to the appropriateness of the factors, which is needed in *justifications* [179].

In our quantitative study, we did not find a main effect of model type and data provenance on perceived integrity (**RQ4.2.1.**). 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 **(RQ4.2.2.)**. We further discuss this finding in section 5.7.4.

5.7.2. 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 [348]). 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 [75, 346], 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 [75]. Difficulties in presenting algorithmic outputs might lead to overreliance on the AI system [72, 198]. End-users' cognitive biases have also been shown to contribute to overreliance [174]. 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 [14]. This is necessary to prevent human intervention from boiling down to a confirmation mechanism of algorithmic outputs [354]. Explanations [411], cognitive forcing functions [72], or reinforcement learning paradigms [73, 226] have been suggested as potential solutions to AI overreliance. For algorithmic decision-making in the public sector, street-level bureaucrats might be better positioned to apply discretion if they were provided with multidimensional outputs and algorithmic *suggestions* instead of mandated outcomes [356]. Designers should evaluate the utility of those solutions while considering the complex bureaucratic processes street-level bureaucrats face in their everyday practice [445].

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 [7].

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) [14]. 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 [182]. 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 [7]. 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 [301].

   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 [108, 125]. 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)) [134].

5.7.3. 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 [445]. By translating high-level system goals into specific design requirements, system designers encode legislation into software [445]. 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 [131] 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 [354], i.e., human intervention should be framed as a *problem of many hands* [92]. This requires to acknowledge the (partial) shift of discretionary power from decision-making time to design time [445], 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* [14]. There are different ways of "datafying" an action or a person [183]. A preventive approach to human intervention [14] advocates for disclosing and challenging the assumptions underneath design choices (and the rationales that led to those choices [436]). Practitioners need both (1) infrastructure [32] and (2) guidance [36, 109, 274] 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 [300]) that shape machine behaviour [322, 323, 350] and the downstream impact of such choices.

Furthermore, with the proliferation of generative AI systems, AI design pipelines are becoming increasingly modular [35, 92]. 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* [92]). 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 [35, 92].

5.7.4. 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.'s [280] 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 [280] 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 [280] was created to capture perceived trustworthiness (conceptualized through perceptions of ability, benevolence, and integrity) towards *human* decision-makers [280]. Even if not explicitly developed for algorithmic decision-making, using the ABI model [280] 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 [186] 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 at al. [279]. 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 [48, 168, 229, 231, 242, 282, 380] (see section 5.2.3), the need for adapting the ABI model [280] 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 5.3.2). Methodological approaches are needed to systematically identify factors unique to algorithmic decision-making and rigorously validate constructs equivalent to the ABI model [280] across different contexts.

5.7.5. 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 [90]. 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 [365, 437], and captured participants' experience renting properties out as short-term rentals as a control variable (section 5.5.1). We did so to avoid *outcome favorability bias* [268, 419] as it has been done in prior work [23, 365, 381]. 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 [211]. Our study might, therefore, be subject to representativeness limitations [252].

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 im-

pact 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 [90]. 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 [348]. 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 [365] 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) [239] or visual means.

5.8. Chapter Takeaways

In this chapter, we presented a mixed-method study that looks into the effect of varying levels of human intervention on decision subjects' fairness perceptions. Through a combination of qualitative and quantitative methods, we evaluated the effect of decision-maker profile, model type, and data provenance on decision subjects' perceptions of ability, benevolence, and integrity. We found that the decision-maker's profile affects perceptions of ability, and benevolence. We also found that perceived ability and integrity relate positively to fairness perceptions. Mediation analyses indicated that the effect of the decision-maker's profile on fairness perceptions is *mediated* by perceived ability and integrity. Decision subjects additionally highlighted the need to ensure data quality, output correctness, and effective discretion from human controllers. Our findings suggest that human intervention shall not be limited to human controllers making the final decision; it should encompass all humans that play a role in developing and deploying AI systems. In chapter 6, we reflect on how a broader interpretation of human intervention relates to the concept of *contestability by design*.

6

Discussion and Conclusion

In this concluding chapter, we first recall the aim of this dissertation. We then summarize our findings and outline how these findings respond to the formulated research questions. We also reflect on the implications of the work included in this dissertation. Next, we point out the limitations of our work. We finalize with some concluding remarks.

6.1. Recalling the Aim

In view of the potentially harmful effects of adopting Artificial Intelligence (AI) systems for decision-making, several actors such as the European High-Level Expert Group appointed by the European Commission have suggested ethical principles for ensuring that such AI systems are trustworthy [140]. One prominent ethical principle for trustworthy AI is *contestability*; a property that makes AI systems open and responsive to human intervention throughout their lifecycles [9]. Contestability has been claimed to be an effective way of empowering decision subjects and of ensuring that algorithmic decision-making processes are regarded as legitimate and acceptable alternatives to human-led decision-making [95, 162, 253, 259, 393, 400].

Despite recent interest, most guidelines for designing contestable AI are conceptual and there is little insight into whether and how contestability empowers decision subjects in algorithmic decision-making. In this dissertation, we addressed this gap by *generating empirical insights into decision subjects' needs for and fairness perceptions towards contestability.* These insights are aimed at informing the development and deployment of contestable AI systems that empower those subjects to algorithmic decision-making. To this end, in chapter 2, we first contextualized contestability in the current discourse on trustworthy AI. In chapter 3, through interview-based qualitative methods, we identified decision subjects' procedural and information needs for meaningful contestability. In chapter 4, we operationalized contestability as the right to contest automated decisions and related safeguards (i.e., explanations, human intervention) as interpreted from Article 22(3) of the European Union's General Data Protection Regulation (GDPR) [131]. We conducted a large-scale quantitative study and evaluated the individual and combined effects of explanations, human oversight, and appeal mechanisms on decision subjects' fairness perceptions. In chapter 5, we further explored how to articulate human intervention as a safeguard that conditions the existence of the right to contest automated decisions as interpreted from Article 22(3) of the GDPR [131]. Through a combination of qualitative and quantitative approaches, we captured decision subjects' fairness perceptions towards algorithmic decision-making configurations with varying levels of human intervention.

6.2. Summary of Findings

RQ1: What are the main trustworthy Artificial Intelligence principles and how are these principles operationalized so as to enable a multi-stakeholder deliberation in AI design and assessment?

With research question 1, we aimed at (1) generating an overview of the current discourse on trustworthy AI so as to locate *contestability* in such discourse. We also aimed at (2) operationalizing trustworthy AI principles and (3) at encouraging a multi-stakeholder deliberation to deal with tensions inherent in the design and assessment of socio-technical systems. To this end, in chapter 2, we conducted a meta-review on recent trustworthy AI standards. The meta-review resulted in the development of a framework that summarizes prominent ethical principles for trustworthy AI design and assessment in a circular arrangement (see figure 6.1). The circular framework presents two bipolar dimensions. Principles that are located close to each other tend to uphold similar values. Opposing dimensions denote tensions between the principles located in those dimensions. The circularity of the framework, therefore, allows to visually identify potential similarities and tensions between different principles. In our framework, contestability is located next to explainability and human control, within the dimension of individual empowerment. Contestability, therefore, upholds values that are closest to explainability and human control, and contributes to the safeguard decision subjects' interests.

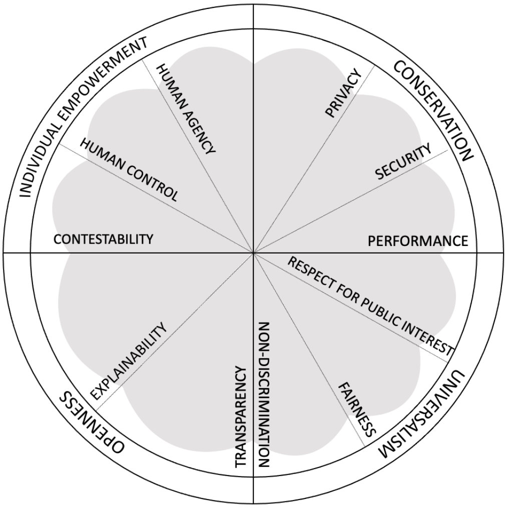


Figure 6.1: Graphic representation of the developed framework.

We operationalized each of the identified prominent principles by breaking them down into criteria that define those principles and their manifestations (see figure 6.2). Since several of these principles and criteria are mutually exclusive and require negotiation, we advocated for engaging several stakeholders in such negotiations. To enable a multi-stakeholder deliberation, we mapped currently available stakeholder-specific means for communicating manifestations of trustworthy AI principles and corresponding criteria. Through this mapping we identified a scarcity of guidelines for operationalizing contestability and means for enabling multi-stakeholder deliberation about contestability.

RQ2: What are decision subjects' needs for meaningful contestability in algorithmic decision-making?

In response to **RQ1**, we identified that, despite contestability being part of the "normative core" [138] that guides a principle-based design of trustworthy AI, the operational-

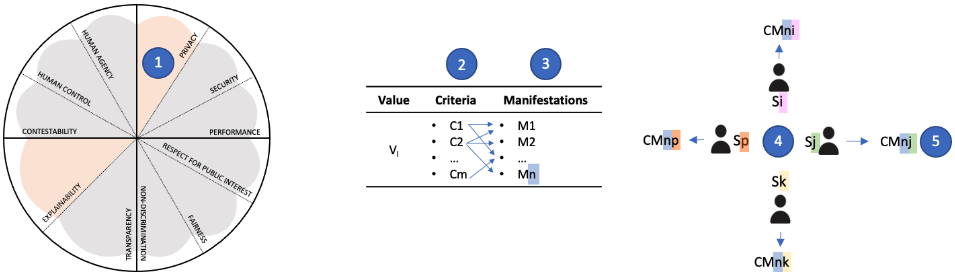


Figure 6.2: Workflow for operationalizing trustworthy AI principles and enabling a multi-stakeholder deliberation. Each value was broken down into specific criteria and manifestations. We then mapped the communication means per stakeholder that were available at the time to deliberate each of the presented values.

ization of and deliberation about contestability had received little attention. Among the different stakeholders that can exercise contestability [9], we decided to further explore decision subjects' needs for contestability. With research question 2, we aimed at generating empirical insights into decision subjects' information and procedural needs for meaningful contestability in algorithmic decision-making. Our findings highlight the cooperative nature of contestability (see table 6.1): decision subjects need (1) cooperation in sense-making, (2) support during contestation acts, and (3) appropriate responsibility attribution throughout the AI development and deployment pipelines.

Cooperation in sense-making. The sense-making process that precedes a contestation act requires the cooperation of decision subjects, legal and AI experts that decision subjects might contacts, and human controllers. The information needs and support level required by decision subjects depended on participants AI literacy and experience with topics related to AI fairness. All participants expressed the willingness to know the *why* behind the decision output. They needed this information to be able to understand their individual situation and to decide whether to engage in an act of contestation. Some decision subjects were willing to know *how* the decision-making process took place. The usefulness of knowing *how* the decision-making took place depended on decision subjects' capacity to make sense of such information. Our results suggest that such sense-making capacity might be mediated by decision subjects' AI literacy. The capacity of using information about the workings of the AI system as part of a contestation process might, instead, be mediated by decision subjects' experience with AI fairness. Regardless of decision subjects' profile, decision subjects requested means to facilitate dialogue with controllers. These included effective communication means and accessible explanations.

Social support in contestation acts. The contestation procedure requires the cooperation between decision subjects, reviewers, third parties assigned to decision subjects, and fellow sufferers. Decision subjects sought organizational and peer support. As far as organizational support is concerned, decision subjects required reviewers of the algorithmic decisions to be flexible, cooperative, empathetic, and experts in AI. Decision subjects additionally required an independent third party to mediate the conflict to compensate power and knowledge differentials. Participants additionally expressed their wish to be part of a collective. This was especially the case when contemplating the possibility of contesting aspects of the AI system itself.

Distributed responsibility. The need for appropriate responsibility attribution requires cooperation between controllers, policy makers and other member of the public administration. Distributing responsibility was highlighted as a means for ensuring algorithmic accountability. Finally, decision subjects required public administration to ensure social transparency (i.e., visibility of interactions within the complex socio-organizational context where algorithmic decision-making takes place) in administrative processes.

These findings highlight the need build capacity for contestation processes, personalize explanations for contestability, and open up sites for contestation throughout AI pipelines.

Table 6.1: Overview of our findings for **RQ2**.

Information and Procedural Needs
T1. Cooperation in Sense-Making – post-hoc intervention
T1.1. Strategizing Information Requests
T1.2. Facilitating Dialogue with Controllers
T2. Social Support in Contestation Acts – post-hoc intervention
T2.1. Seeking For Organizational Support
T2.2. Seeking For Peer Support
T3. Distributed Responsibility – ex-ante intervention
T3.1. Ensuring Algorithmic Accountability
T3.2. Fostering Social Transparency

RQ3: How do elements related to "the right to contest" automated decisions as interpreted from the European Union's General Data Protection Regulation (i.e., explanations, human intervention, appeal mechanisms) affect decision subjects' fairness perceptions?

With research question 3, we aimed at generating empirical insights into decision subjects' informational and procedural fairness perceptions towards contestable algorithmic decision-making configurations. To this end, we brought in perspectives of legal scholars on contestability to complement the discourse in the field human-computer interaction. We narrowed down contestability to appeal mechanisms. We operationalized contestability as the *right to contest* automated decisions and related safeguards (i.e., explanations, human intervention) as interpreted from Article 22(3) of the General Data Protection Regulation [131]. We found that (1) explanations and (2) appeal mechanisms positively impact decision subjects' informational and procedural fairness perceptions, respectively; we did not find find an effect of (3) human oversight on decision subjects' fairness perceptions; we found that (4) informational and procedural fairness perceptions positively relate to overall fairness perceptions (see Figure 6.3).

Explanations. Our findings showed that the presence of explanations positively contributes to informational fairness perceptions. Exploratory results additionally pointed

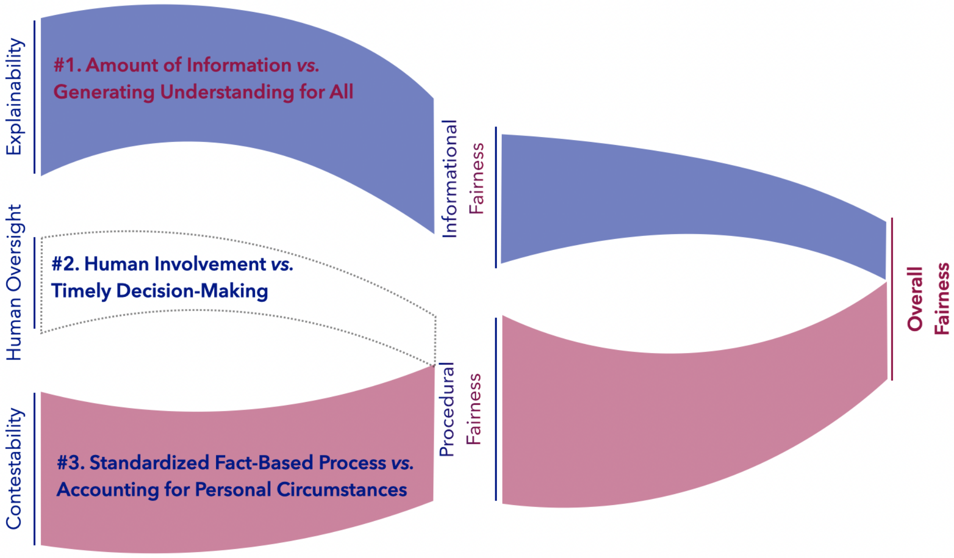

Explainability
#1. Amount of Information vs. Generating Understanding for All
Human Oversight
#2. Human Involvement vs. Timely Decision-Making
Contestability
#3. Standardized Fact-Based Process vs. Accounting for Personal Circumstances
Informational Fairness
Procedural Fairness
Overall Fairness

Figure 6.3: Summary of our findings in response to **RQ3**.

towards an interaction effect between the presence of explanations and AI literacy. Responses to open-ended questions surfaced a tension between the optimal amount of information to be provided and enabling understanding for all.

Contestability — operationalized as appeal mechanisms. Our results also showed that the presence of contestability, operationalized as appeal mechanisms, positively contributes to procedural fairness perceptions. Despite the positive effect of appeal mechanisms on decision subjects' procedural fairness perceptions, the presented appeal processes were still negatively perceived in procedural voice and outcome influence on a −3 to +3 scale. The option to contest the initial decision scored ($M = -0.81, SD = 0.17$) and ($M = -1.30, SD = 0.16$) on procedural voice and outcome influence, respectively. The option to contest the decision-maker scored ($M = -1.21, SD = 0.16$) and ($M = -0.65, SD = 0.19$) on procedural voice and outcome influence, respectively. Responses to open-ended questions surfaced a tension between enabling contestability to account for personal circumstances and ensuring a standardized fact-based process.

Human oversight. We did not find an effect of human intervention (in the form of human oversight) on decision subjects' procedural fairness perceptions. Responses to open-ended questions indicate that decision subjects required a higher level of human involvement in the algorithmic decision-making process. However, this conflicted with their willingness to be subject to a timely decision-making as involving humans in the loop would slow down the decision-making process.

Multi-dimensionality of fairness perceptions. Our results also showed that both informational and procedural fairness perceptions positively contribute to overall fairness perceptions; with procedural fairness perceptions having a bigger impact. We did not find an interaction effect between informational and procedural fairness perceptions.

These findings highlight the need to leverage transparency beyond outcome explanations while making sure explanations are accessible for all. The results also encourage to rethink appeal mechanisms so as to give effective voice to decision subjects. The need to further look into effective ways of integrating human intervention in algorithmic decision-making is also evident from our results.

RQ4: How do varying levels of human intervention affect decision subjects' fairness perceptions in algorithmic decision-making?

In response to **RQ3**, we did not find an effect of human intervention (in the form of human oversight) on decision subjects' fairness perceptions. Despite this lack of empirical evidence, current policy efforts heavily rely on human intervention as a safeguard of decision subjects' fundamental rights. According to interpretations of Article 22(3) of the GDPR [131], human intervention shapes the existence of decision subjects' right to contest automated decisions. Furthermore, human intervention might be interpreted as the only safeguard that can prevent a chain of automatisms from happening if contestations are dealt with through algorithms. Given the relevance of human intervention, with research question 4, we aimed at further exploring the safeguard of human intervention and characterizing the effect of human intervention on decision subjects' fairness perceptions. To this end, we first explored models in organizational psychology that would appropriately capture decision subjects' perceptions towards decision-makers that present varying levels of human intervention; we identified the Ability, Benevolence, and Integrity model (ABI model [280]) as an appropriate model to capture decision subjects' perceptions towards decision makers. We also evaluated how decision subjects' perceptions towards decision-makers related to fairness perceptions.

Our qualitative findings indicated that the decision-maker profile might affect the way decision subjects perceive decision-makers' ability and benevolence. Qualitative findings also indicate that model type and data provenance might affect the way in which decision subjects perceive decision-maker's integrity. Interview responses also indicated that decision subjects insist on the need to ensure data quality, output correctness, and the need for decision-makers to effectively apply discretion.

Our quantitative results showed that decision-maker profile affects decision subjects' perceptions of ability, benevolence, and integrity. Perceptions were more positive for hybrid decision-maker configurations than fully-automated ones. Interestingly, benevolence perceptions in all cases were negative. Moreover, we found that the effect of decision-maker profile on decision subjects' fairness perceptions was *mediated* by perceived ability and integrity. We did not find an effect of model type, and data provenance on perceptions of integrity, but, through an exploratory analysis, we found that there was an interaction effect between them. We also found that perceptions of ability and integrity positively contribute to fairness perceptions (see Figure 6.4).

Our qualitative and quantitative findings indicate the multi-faceted nature of human intervention. When interpreting human intervention uniquely as the intervention by the human controller involved at the end of the decision-making process, our findings highlight the central role of administrative discretion in upholding decision subjects' justice standards. When interpreting human intervention as the intervention of several humans throughout the lifecycle of AI systems, our findings highlight the importance of exam-

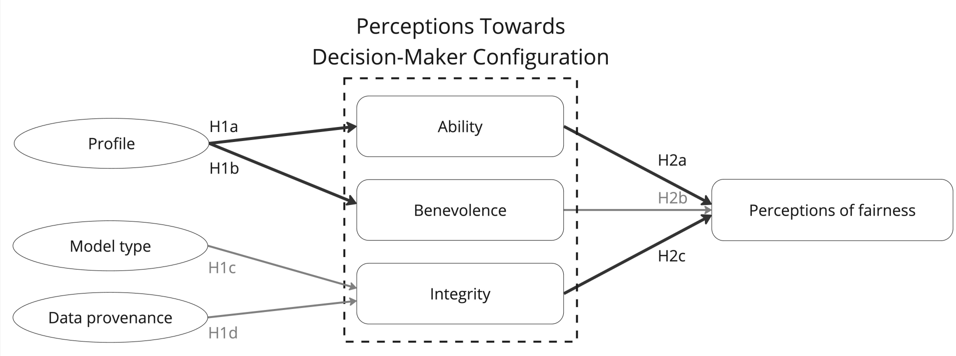


Figure 6.4: Summary of our findings in response to **RQ4**.

ining and documenting practitioners' practices as a way to ensure the integrity of the development process (e.g., data quality and output correctness).

6.3. Implications and Future Work

In this section, we discuss implications of our work along with future research directions stemming from these implications.

6.3.1. Empowering Decision Subjects to Exercise Contestability

The findings in chapter 4 showed the positive effect of contestability on decision subjects' fairness perceptions. Yet, our results suggest that directly transferring appeal mechanisms designed for human-led decision-making to algorithmic decision-making does not address decision subjects' needs for contestability. This means that contestability for algorithmic decision-making needs rethinking some aspects of contestations that are unique to algorithmic decisions. Rethinking contestability involves several challenges and future research opportunities.

Promoting understanding. Our findings in chapter 3 suggest that understanding an algorithmic decision poses additional challenges compared to understanding a decision that is uniquely based on human-led processes. Decision subjects need to make sense, understand and strategically use information about the algorithmic decision-making as part of a contestation. This sense-making process depends on different factors, among which we identified two prominent ones: AI literacy and experience with fairness in AI. Our results in chapter 4 additionally showed that AI literacy interacts with explanations when predicting decision subjects' informational fairness perceptions. A direct implication of our findings is that there is no one-size-fits-all solution to providing functional and actionable explanations for contestability. Depending on individual factors, an explanation —not necessarily limited to the algorithmic output— might be actionable or not for a decision subject. Two interrelated future research opportunities emerge from this phenomenon: (1) the need to characterize sense-making processes for contestability, and (2) the need to design explanations that generate understanding among diverse individuals. The first line of research involves conducting observational studies and engaging with theoretical frameworks coming from disciplines like information processing or social cognition. This will generate insights into how decision subjects (individually or collectively) make sense of explanations and might help identify additional factors that affect such process. Characterizing sense-making processes can then inform the design of explanations that are adaptable to different individuals' needs. Different amounts of information, granularity, or explanation modalities are just a few of the parameters that could be tweaked to iterate on the design of actionable explanations for contestability [101, 235, 431].

Building capacity for contestability. Beyond the differences in information needs, our findings in chapter 3 show that the processes needed to handle contestations in algorithmic decision-making also present some additional challenges compared to human-led decision-making. An example is that decision subjects required reviewers to be knowledgeable in AI and able to translate information about algorithmic decisions into accessible explanations. Additionally, decision subjects expressed the need to involve a third party to mediate the conflict. This third party should (1) provide similar examples to decision subjects' individual situations and (2) identify whether the AI might have made a potential error. Such requirements would demand considerable effort from an organization to train their workforce in AI; decision-makers are normally experts in the topic they deal with, but not necessarily experts in AI. It would also require organizations to build capacity to closely monitor the performance of the AI. Our work, therefore, implies that, when switching to algorithmic decision-making, organizations should not only consider the resources needed to adopt and monitor the main algorithmic decision-making process, but also the additional costs that erroneous algorithmic decisions might incur [333]. This is especially relevant in the case of the public sector, where decisions are made in a context of resource scarcity and bureaucratic burden [356].

Enabling collective action. Our results in chapter 3 showed that the idea of individually contesting the usage, design, or integration of AI systems in decision-making processes was embraced by very few decision subjects. These had high AI literacy and extensive experience with AI fairness. Instead, if found problematic, the option to collectively contest the adoption of AI systems for decision-making was mentioned by many. Decision subjects regarded collective action as a means for ensuring that their "rights, freedoms, and legitimate interests" [131] are safeguarded. They regarded collectives as effective consortia that could identify patterns of erratic algorithmic behavior and help individuals remedy their situation. Enabling collective action brings in challenges that future research should look into. We will highlight three: (1) accounting for decision subjects' imaginaries, (2) scaffolding collective action, and (3) incentivizing collective action. First, for decision subjects to engage in collective action, they need to understand that AI systems might lead to potential errors or risks. This requires raising awareness of the fact that AI systems are not always "100% right and safe" [211]. Public entities advocating for contestability in algorithmic decision-making should, therefore, look into ways that could help lay citizens acquire an intuition of what an AI system can and cannot do. Further research is needed to identify the best communication means (e.g., videos [406], comics [418, 420] or stories [406]) for this purpose. Second, support structures are needed to scaffold and formulate collective action, when appropriate. Previous work has suggested different formats for collective action, both in a post-hoc and ex-ante fashion. Collective action can take place during the development and deployment of AI if decision subjects are involved early in the design process. Participatory frameworks like *WeBuildAI* [250] explore computational models that represent decision subjects' views in policy creation. This enables design choices to be collectively agreed or contested. When it comes to post-hoc collective action, the *Contestation Café* [93] has been suggested as a speculative concept where a panel of "Fixers" –people with the right expertise– help decision subjects identify and challenge unfair algorithmic outcomes. Future research could use these two examples as starting points to structure effective means for collective action. Third, collective action requires participation, and participation requires incentives. Initiatives for engaging decision subjects in open debates about potential problematic aspects of using AI systems can be of great benefit. If the decision-making organization (e.g., public administration) is responsible for incentivizing the corresponding support structures, economic and/or political interests should not derail the impartiality and independence of support structures that should be at the service of decision subjects.

Normative considerations. Contestability in algorithmic decision-making is a relatively new research area, and still lacks clear normative guidelines on how contestability should look in practice. While the GDPR [131] defines decision subjects' *right to contest* automated decisions, it does not specify when or how decision subjects can oppose an automated decision. Our work provides insights into what decision subjects' would ideally require to meaningfully contest algorithmic decisions. However, policy decisions about what can be contested, who can contest automated decisions, who is accountable, or the type of reviews that should be in place are yet to be made [269]. Policy efforts in this direction will further help contextualize and ground the insights we generated through our work. This will help design for contestability with decision subjects' empowerment in mind, while being consistent with legislation.

The burden of contestability. Our work showed the potential of contestability as a prominent ethical principle for empowering decision subjects, but it also acknowledges the burden that contestability might put on those subject to algorithmic decisions. From our findings in chapters 3 and 4, it can be seen that being able to contest automated decisions is viewed more positively by decision subjects than not being able to contest these. However, decision subjects would prefer not having to deal with contestation processes themselves if algorithmic decisions are mistaken. A direct implication of these findings is that contestability, as exercised by decision subjects, should not be seen as an easy shortcut to comply with legislation, yet disregarding quality control [160]. The algorithmic imprint (i.e., long-lasting harmful consequences of algorithmic systems [128]) of erroneous algorithmic outputs cannot be ignored, even if these erratic systems are contested and eventually removed. While decision subjects should be able to exercise their right to contest algorithmic decisions, when erratic outputs are concerned, contestability might be better exercised by other stakeholders, e.g., practitioners. Enabling practitioners to exercise contestability requires further research into characterizing workflows for the identification of potential erroneous algorithmic outputs and for the improvement of AI robustness. For example, more studies are needed to characterize practices for debugging models and ensuring robustness [395]. Developing tools that can support practitioners in debugging processes is also an important future research direction [34].

Inspecting these practices and related infrastructure contributes to the quality assurance plan for contestable AI systems [9].

6.3.2. Broadening the Understanding of Human Intervention

The results from chapter 5 suggest a positive effect of human intervention on decision subjects' fairness perceptions. While including a human in the loop as part of algorithmic decision-making was positively perceived, decision subjects did emphasize the need for human controllers to apply their tacit knowledge and exercise discretion. Decision subjects required human controllers to have the ability to evaluate their individual circumstances and influence the output of the AI system. Our findings have several implications for the field of human-computer interaction.

Human intervention as a means to discretion. The field of human-computer interaction (HCI) has long been aware of the impact that cognitive biases might have on the way users interact with AI systems. Cognitive biases might impact human controllers' ability to discern cases that need accounting for decision subjects' individual circumstances and might lead to overreliance [72, 174, 198]. Several solutions have been suggested to ensure human controllers spot potentially erroneous outputs of the system and appropriately rely on AI. One of the most widely studied solutions to overreliance has been the generation of explanations about algorithmic outputs (e.g., [246, 363, 411]). A recent study has shown that providing explanations *can* be an effective way of reducing overreliance if explanations reduce the costs of verifying the AI's prediction; it is in this case that human controllers are willing to engage with explanations [411]. Other strategies include cognitive forcing functions [72] (e.g., slowing down the decision-making process), which are claimed to be more effective than simple explainable AI approaches in reducing overreliance. For the specific context of public algorithmic decision-making, scholars have also advocated for making algorithmic outputs multidimensional [356]. Multidimensionality allows human controllers to have flexibility when interpreting algorithmic outputs and to exercise discretion. Our findings reinforce the idea that hybrid decision-making configurations are preferred to fully automated ones and, therefore, encourage further HCI research in understanding cognitive biases and suggesting context-specific solutions to reduce overreliance. This should include research into e.g., human factors that contribute to overreliance, context-specific preferences towards specific solutions that reduce overreliance, interfaces that allow human controllers to exercise discretion while adapting to existing workflows in real-world contexts. Deepening our knowledge in this area will ensure that human intervention is, indeed, a valid safeguard to uphold the interests and rights of decision subjects.

Understanding human intervention as a problem of many hands. In our work, we have limited human intervention to human controllers who can change algorithmic outputs and are involved at the very end of algorithmic decision-making processes. However, humans can intervene at many different points across the pipelines that compose AI systems [14]. Our findings highlight that, even if not mentioned explicitly, decision subjects required human intervention to be understood in a broad sense. For example, when decision subjects requested to ensure data quality (chapter 5), they were indirectly pointing to the need to inspect the "undisclosed yet impactful subjective choices" [78] that humans involved throughout AI pipelines make in their everyday practices. Data quality is impacted by human-led decisions about which data to include or exclude when training AI systems. These decisions might end up injecting representation biases [387] or generating syntactic data silences (i.e., un-inclusion of sub-populations) [300]. When human intervention is understood as a *problem of many hands* [92], advocating for effective human intervention requires raising awareness about the discretionary choices made during the development and deployment of AI systems. It also requires researching into several individuals, their practices and tools involved throughout AI pipelines so that these choices become transparent and open to dispute. This is linked to the interpretation of the right to contest in Article 22(3) of the GDPR [131] as the apex of contestability as an architectural principle of AI systems (i.e., contestability by design) [355]. We devote the full subsection 6.3.3 to describing the challenges and future research directions in the field of human-computer interaction for upholding contestability as a design principle.

6.3.3. Ensuring Contestability By Design for AI Systems

In order to ensure that AI systems used for decision-making are open to scrutiny and dispute throughout their lifecycles (i.e., they are contestable by design [9]), transparency needs to become a norm in different stakeholders' daily practices —see Figure 6.5. Given the complex and distributed nature of the human labor that AI systems result from, opening up sites for contestation throughout AI systems lifecycles presents several challenges. We believe HCI scholars are well positioned to contribute in this research direction.

Promoting reflexivity. Oftentimes, technical work that leads to the development and deployment of AI systems is viewed as "objective". However, there are various value-laden choices that different stakeholders make throughout AI systems lifecycles [322, 323, 350]. For example, during the very early stages of an AI project, business goals are translated into design requirements. This means that appropriate target variables or proxies to these variables are defined [321]. When dealing with data, practitioners determine which data is included or excluded, and what is considered *good* data [328]. All these choices are value-laden. In a similar vein, decisions about how to transform, analyze, and visualize data are also far from being "objective" decisions. These decisions are rarely disclosed yet considerably impact the behavior that AI systems display downstream [78]. Making discretionary choices salient can help increase transparency and accountability in AI systems. The first challenge of making discretionary choices salient involves identifying, reflecting, and acknowledging the value-laden nature of such choices. HCI scholars can contribute in this direction by (1) developing adequate infrastructure that promotes reflexivity, and by (2) further training and guiding practitioners in reflecting upon their everyday practices. When it comes to technical infrastructure for promoting reflection, the technique suggested by Cambo et al. [78] is a good example. Cambo et al. [78] suggested a technique to digitally exhibit data annotators' behaviors and visualize similarities and differences among annotators. The aim of this technique is to provide data scientists a representation of the annotators' positions with respect to the annotated data. This allows data scientists to reflect about the social impact of their models in advance. Beyond data annotation, further research is also needed to develop technical interfaces that enable data versioning and to keep track of the effects of data

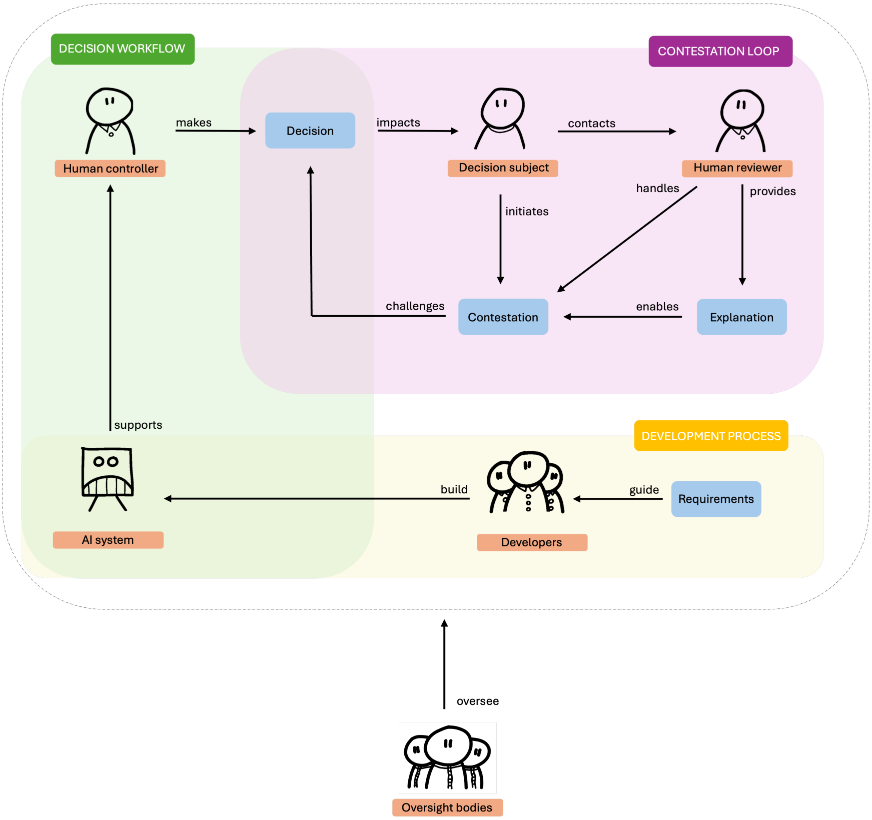


Figure 6.5: Actors and processes that are part of contestable AI (adapted from [7, 9]). While the focus of this dissertation was on understanding *decision subjects'* needs for and perceptions towards contestability, contestable AI requires opening sites of contestation throughout the lifecycle of AI systems. This includes, not only actors involved in the contestation loop, but also actors, like developers or human controllers, involved in the development process and the decision workflow.

transformations on algorithmic fairness [32, 359]. As far as practitioners' training is concerned, prior work in the usage of fairness toolkits has highlighted the need to provide practitioners with real-world examples and onboarding materials that offer guidance for reflecting on the consequences of their choices [109, 188, 244]. Practitioners should also be warned against turning reflexive practices in a mere checkbox culture [36].

Opening up sites for contestation. Many of the value-laden decisions that lead to the development and deployment of AI systems are not a given, but are rather negotiated at different stages of AI pipelines [436]. Negotiations that happen at critical junctures of AI pipelines (e.g., at the problem formulation stage) should open up spaces for contestation and collective reflection. These negotiations should bring together different stakeholders with backgrounds in both technical and non-technical disciplines. As shown in chapter 2 of this dissertation, enabling fruitful and informed negotiations among stakeholders is not a trivial task. Human-computer interaction scholars can contribute to opening up sites for contestation by (1) inspecting current workflows and practices during the implementation of AI projects and by (2) identifying relevant junctures in which value-laden decisions should be open to scrutiny and dispute. While this dissertation does not include contributions in this direction, we are currently exploring decision-making dynamics during AI projects both in the Dutch public sector and several European private companies. Characterizing such dynamics allows us to open up sites for contestation throughout AI pipelines while accounting for existing workflows. Additional HCI contributions for fostering collective reflection would involve the development of tools that facilitate communication among stakeholders [109, 188, 335] and digital objects (e.g., data visualization techniques) that enable fruitful negotiations [184].

Accounting for algorithmic supply chains. From the above reasoning, one could argue that we are limiting future challenges and research opportunities in contestability to organizations that develop AI systems in-house. However, with the proliferation of Generative AI (GenAI), algorithmic systems are increasingly modular and result from complex supply chains. In algorithmic supply chains, the development and deployment of AI systems is distributed between interdependent actors. Responsibility is also distributed across those actors. Due to such distribution of tasks and responsibilities, organizations developing AI systems have little visibility and understanding of the context of deployment. Similarly, organizations deploying AI systems have little influence over how those systems are developed [92]. This leads to a situation where different actors suffer from *accountability horizon* (i.e., point beyond which an actor has no visibility of the choices made in the supply chain) [92]. Accountability horizon makes it difficult to ensure transparency and, consequently, to uphold the principle of contestability by design. Accounting for algorithmic supply chains when designing for contestability is a key aspect that future research should consider. This leads to additional challenges and research opportunities. On a policy level, legal and institutional mechanisms are needed to foster visibility across algorithmic supply chains and to ensure that design choices are open to dispute [92]. Future policy efforts should, therefore, account for the dynamics of algorithmic supply chains and frame their risk management approaches accordingly. When it comes to HCI scholars, future research should study various aspects of supply chains, e.g, who is involved in algorithmic supply chains, how these chains are structured, how supply chains develop over time, or how decision-making processes take place across supply chains [92]. Bringing transparency to the dynamics governing algorithmic supply chains represents a first step towards effectively designing for contestability in a context of distributed yet interdependent responsibilities. While not part of this dissertation, we are currently exploring what contestability looks like in algorithmic supply chains. We have organized a workshop at the 2024 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) that explores the challenges of ensuring contestability in algorithmic supply chains [33]. This workshop brought together HCI scholars to collectively reflect on the challenges mentioned above and identify additional ones.

6.3.4. Recommendations Per Stakeholder

In this subsection we summarize the implications outlined in sections 6.3.1 to 6.3.3 in a set of recommendations targeted at different stakeholders.

- **Recommendations for Organizations Developing and Deploying AI systems.**
  1. Tailor explanations about algorithmic outputs and decision-making processes to decision subjects with different levels of AI literacy.
  2. Build capacity to ensure decision subjects feel empowered in contestation processes. For instance, by implementing effective communication channels, by ensuring reviewers have relevant AI knowledge, or by providing decision subjects access to a third party that can mediate the conflict.
  3. Enable collective contestations. Arrange support structures that help decision subjects formulate collective contestations. These support structures should be aligned with the principle of impartiality.
  4. Ensure robustness of AI systems and correctness of algorithmic outputs. Avoid employing contestability, exercised by decision subjects, as a safeguard against erratic AI systems.
  5. Provide human controllers with relevant metrics and interfaces to effectively exercise discretion.
  6. Balance the need to rely on models that provide justifications about the decision and decision subjects' right for privacy.
  7. Disentangle decision subjects' fairness perceptions towards algorithmic decision-maker configurations and fairness perceptions towards the implemented policy.
  8. Train and guide practitioners in incorporating reflection in their everyday practices. Build teams with technical and non-technical expertise.
  9. Bring different stakeholders together in negotiation processes. Foster collective reflection in key decision-making junctures along AI projects.

- **Recommendations for Policy Makers.**
  1. Promote policy efforts to define normative boundaries for contestability. Define what can be contested, who can contest algorithmic decisions, how reviews should take place and who is accountable.
  2. When defining human intervention, account for the complex and distributed human labor that AI systems result from. Frame human intervention as a *problem of many hands.*
  3. Foster visibility across supply chains. Account for dynamics in algorithmic supply chains when framing risk management approaches for AI systems.

- **Recommendations for HCI Scholars (future research directions).**
  1. Characterize sense-making processes that precede contestations. Characterize factors that impact such sense-making processes and the intersection of those factors.
  2. Explore the design of explanations that are adapted to the needs of different decision subjects and that generate effective understanding for contestability.
  3. Inspect and characterize dynamics of participation and incentivization in collective contestation processes.

4. Characterize practices for debugging models and develop tools that support practitioners in ensuring robustness of AI systems.
5. Characterize cognitive biases that lead to overreliance on AI systems and possible solutions to overcome these cognitive biases.
6. Characterize human controllers' needs to effectively exercise discretion and factors that affect these needs in real-world contexts.
7. Inspect ways in which human intervention is being shaped in real-world contexts and whether or how decision subjects' fundamental rights are considered as a central aim.
8. Explore which technical infrastructure is effective in promoting reflection among AI practitioners.
9. Characterize current workflows and negotiation practices in AI projects. Identify relevant value-laden decisions that should be open to scrutiny and dispute at different junctures of AI pipelines.
10. Develop tools to facilitate collective reflection in decision-making junctures.
11. Characterize dynamics in algorithmic supply chains. Identify e.g., who is involved, how supply chains develop over time, how decision-making processes take place in algorithmic supply chains. Determine how these dynamics affect transparency and contestability by design.

6.3.5. Reflections on Our Approach

In this dissertation, we captured decision subjects' needs for and fairness perceptions towards contestability. To conduct our work we relied on a mixed-methods approach, grounded in two different decision-making contexts: one in the public sector and the other one in the private sector. In this section, we reflect on the effectiveness of our approach in generating new knowledge on how contestable AI systems should be designed and deployed.

Capturing Decision Subjects' Needs and Perceptions to Inform the Development and Deployment of Contestable AI. Ours is part of a broader effort to ensure that the development and deployment of AI systems is conducted in a human-centered way [374]. The ultimate goal of human-centered approaches to AI is to avoid harmful and undesired effects of AI systems [377]. Our work has relied on a combination of interviews and user studies to capture decision subjects needs' and fairness perceptions towards contestability. This combination has allowed us to (1) generate nuanced insights into decision subjects' needs for contestability, and to (2) characterize the more basic perceptions that drive those needs. In both our interview and user studies, we have relied on indirect ways of capturing decision subjects' needs for and perceptions towards contestability. Future research could additionally engage with participatory approaches that directly involve decision subjects in AI design processes by following e.g., the participatory workshops run by Vaccaro et al. [407] to inform the design of contestable content moderation or the workshops conducted by Brown et al. [70] to capture the perspectives of affected communities towards the use of AI systems in child welfare services. When it comes to evaluating contestable AI systems, a promising approach is that of end-user audits [108, 112, 238, 372]. In end-user audits, the lived experiences of everyday users are leveraged through crowdsourcing. These experiences help uncover harmful behaviors of AI systems. While this approach has not been used as part of this dissertation, we have organized a workshop at the 12th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024) to explore the potential of crowdsourcing for engaging end users in testing, auditing and contesting AI systems. This workshop brought together HCI and AI scholars (focused on human computation) to explore the design and deployment of crowdsourcing pipelines that enable responsible AI auditing and evaluation. We reflected on topics such as (a) future mechanisms for scaffolding crowds in auditing and evaluation processes, (b) strategies for ensuring crowd diversity, or (3) psychological aspects generative AI auditing and evaluation through crowdsourcing.

Combining qualitative and quantitative approaches. Through a combination of qualitative and quantitative approaches, we have been able to generate insights into decision subjects' needs for, and perceptions towards, contestability. Qualitative studies allowed us to acquire a nuanced understanding of decision subjects' needs for contestability. In our qualitative studies, we involved short-term renters, i.e., participants who had something at stake in the studied context. Their lived experience and first-hand testimonies brought perspectives that we could have not gained ourselves. These perspectives also allowed us to narrow down the research questions that we would ask in our quantitative studies. Through qualitative studies, we also identified potentially useful theoretical models to capture decision subjects' perceptions towards decision-makers (i.e., the ABI model [280]). We then modeled the nuanced insights we gained through qualitative studies for these to be tractable through quantitative approaches. Quantitative studies allowed us to test whether the perceptions that we identified in our small pool of interviewees were also applicable to a bigger pool of participants.

Differences and commonalities between the public and private sectors. In section 1.4.3, we pointed three differences between the private and public sectors (i.e., lack of alternatives, need for administrative discretion, and societally sensitive topics that the public sector deals with, unlike the private sector) and motivated why we looked into both sectors. Differences between the public and private sectors manifested in our findings. First, in the private sector, decision subjects highlighted the positive impact of adopting AI systems for loan approvals, even if they received a negative decision. Time efficiency was claimed to be a positive aspect of algorithmic decision-making since it would give decision subjects a rapid response and an opportunity to look for an alternative in another financial entity. This was not the case in the public sector. Decision subjects have no alternative in the public sector and, therefore, time efficiency would not make up for incorrect or unfavorable algorithmic outputs. However, in the public sector, some decision subjects *did* highlight the benefit of accelerating policy enforcement decisions as a way to ensure a greater social good (i.e., in the studied context, access to housing). Second, both in the public and private sectors decision subjects asked human controllers to attend to individual circumstances when incorporating AI systems in their workflow, yet with different motivations. In the private sector, attending to individual circumstances was linked to wishes of compassion and empathy from the human controller. Instead, attending to individual circumstances, or, in other words, exercising discretion, was linked to the very nature of the decision-making process in the public sector. Third, because of the nature of the explored public decision-making context (i.e., policy enforcement), decision subjects requested justifications rather than explanations. This was not the case in the private sector, where the decision-making process involved a prediction problem rather than a policy enforcement one.

All in all, inspecting the private and public sectors has shed some light on how some of our findings apply to both sectors, while others are either shaped by the nature of the particular decision-making problem or unique to the context. Such an observation should act as a deterrent against suggesting one-size-fits-all solutions to contestability. It should rather encourage HCI scholars to inspect and acknowledge the peculiarities of the contexts that we study.

6.4. Limitations

This section aims at acknowledging the limitations of the work presented as part of this dissertation.

Participant recruitment and incentivization. The cultural and geographical background of our participants might have influenced the results presented in this dissertation. For the qualitative studies presented in chapter 3 and chapter 5 (study 1), participants were recruited by publishing the call for participation in online housing channels, by putting posters around TU Delft and by reaching out to personal contacts. Participants were then selected so as to obtain a pool of individuals with varying levels of AI literacy. All participants were educated individuals with high digital literacy. For the quantitative studies presented in chapters 4 and 5 (study 2), instead, we used crowdsourcing platforms to recruit participants. All crowdworkers were coming from the Global North and were proficient in English. Previous work [28, 54, 211] has shown that decision subjects' backgrounds affect their fairness perceptions towards decision-making processes. In a similar vein, our participants received monetary incentives to participate in our studies. This might have impacted our research outcomes. Future research should further explore whether our results hold when participants coming from the Global South are recruited or when participants are incentivized with rewards other than monetary ones.

Vignettes as data generation method. The usage of vignettes as prompts to generate data about decision subjects' needs for and perceptions towards contestable algorithmic decision-making configurations might have impacted the way our research process and outcomes unfolded. Individuals that participated in our studies either had a stake in the context at hand —qualitative studies presented in chapters 3 and 5 (study1)— or were asked to empathize with decision subjects' position —quantitative studies in chapters 4 and 5 (study 2)—. Vignettes were used in both qualitative and quantitative studies for participants' needs and perceptions towards the presented scenarios to manifest. Participants were shown the corresponding vignettes once. Future research should look into how different interaction modalities, data generation methods, or longer research timescales might lead to different or additional insights.

Transferability of results to other contexts. The results obtained through the empirical studies included in this dissertation are limited by the nature of the explored decision-making contexts. The studies included in this dissertation were grounded in two different decision-making contexts; one in the public sector (chapters 3 and 5) and the other one in the private sector (chapter 4). The decision-making context in the public sector consisted of an algorithmic illegal holiday rental identification scenario. The decision-making context in the private sector consisted of an algorithmic loan approval scenario. When it comes to decision-making, the public and private sectors are distinct (see section 1.4). Grounding our research in both sectors shed some light on how our results might be transferable from one sector to the other. However, there are several other use cases that we could have explored both in the public and private sectors. Future work should explore additional use cases in order to determine the extent to which our results are transferable to such contexts. An example in this direction includes a recent study by Aljuneidi et al. [11] . Aljuneidi et al. [11] showed that, for a hypothetical algorithmic identity card renovation scenario, the presence and level of detail of explanations affect decision subjects informational fairness perceptions. These results highly resonate with the results presented in chapter 4 of this dissertation and suggest that our results might be transferable to additional contexts.

6.5. Concluding Remarks

This dissertation has contributed to the literature on contestable Artificial Intelligence by generating empirical insights into decision subjects' needs for and fairness perceptions towards contestability in algorithmic decision-making. In chapter 2, we identified *contestability* as a prominent ethical principle for trustworthy AI, yet with a scarcity of guidelines for its operationalization or means for multi-stakeholder deliberation (**RQ1**). In chapter 3, we identified a variety of decision subjects' information and procedural needs for meaningful contestability (**RQ2**). In chapter 4, we identified the need to rethink the design of appeal processes for algorithmic decision-making, as well as the need to further look into human intervention as a key safeguard that conditions decision subjects' right to contest automated decisions (**RQ3**). In chapter 5, we provided insights that suggest that human intervention should not be limited to the human controller making the final decision, but should rather be framed as *a problem of many hands* (**RQ4**).

In sum, empirical insights generated as part of this dissertation emphasize the importance of contestability for empowering decision subjects as part of algorithmic decision making. However, organizations developing and deploying AI systems need to rethink the way in which these contestations take place. Traditional appeal mechanisms place a considerable burden on decision subjects. Our findings suggest that meaningful contestability in algorithmic decision-making requires the cooperative effort between civil servants in several different roles (i.e., street-level bureaucrats acting as controllers, reviewers, policy makers, AI developers and deployers), citizens and third parties (e.g., legal counselors). This dissertation additionally generated empirical insights that highlight and confirm the need to ensure *contestability by design* as a property of AI systems. Contestability by design requires the activity of organizations and individuals involved in the development and deployment of AI systems to be transparent and open to dispute. This should involve research into the practices and perceptions of additional stakeholders who can and should exercise contestability across AI pipelines. The right to contest automated decisions provided by Article 22(3) of the GDPR [131] shall, therefore, not be interpreted an easy shortcut to legal compliance at the expense of algorithmic output correctness. It should rather act as a catalyst for the acknowledgment, reflection and scrutiny of the value-laden discretionary choices that make up algorithmic supply chains. Towards the development of practices and tools that embed contestability as an architectural principle in the design of AI systems.

Appendices

Appendix A: Chapter 2

Appendix A.1.: Tailored communication of system-related information

		Development team	Auditing team	Data Domain experts	Decision subjects
Conservation	Privacy	[K]	[K]		[A] [B]
	Security	[K] [W] [AB]	[K] [W]		
	Performance	[F] [G] [H] [Y] [Z] [AE]	[G] [H] [Y] [Z] [AE]	[I] [J]	[J]
Universalism	Respect for public interest	[E] [AE]	[E] [AE]	[E]	[C] [D]
	Fairness	[G] [H] [K] [W] [X] [Y] [Z] [AD]	[G] [H] [K] [W] [X] [Y] [Z] [AD]	[I] [J]	[J]
	Non-discrimination	[H] [K] [X] [Y] [AD]	[H][K] [X] [Y] [AD]	[J] [L]	[J] [L]
Openness	Transparency	[H] [K] [M]	[H][K] [M]	[I] [J] [L] [M]	[B] [J] [L] [M]
	Explainability	[M] [N] [O] [Q] [AC] [AD] [P]	[M] [N] [O] [Q] [AC] [AD] [P]	[J] [M] [N] [O] [Q] [P]	[J] [M] [N] [O] [Q] [R] [S] [P]
Individual empowerment	Contestability	[U]	[U]	[T] [U]	[T] [AF]
	Human Control	[V]	[V]	[T] [V]	[C] [T] [V]
	Human Agency			[T]	[T] [B] [AA]

Table 6.2: Mapping of available means for transmitting value-specific manifestations to different stakeholders based on the purpose of their insight and the nature of their knowledge. These means have been classified into three main categories: descriptive documents specifying whether/how a value manifestation is fulfilled (red), strategies for fulfilling value manifestations (blue), and complete tools for enabling the fulfillment of value manifestations (green). This table aims at facilitating the navigation of table 6.3, where each means is documented.

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[A]	Iconsets for data privacy declarations [131, 189, 283, 343]	Privacy	• Description of what data is collected • Description of how data is handled • Purpose statement of data collection • Statement of how long the data is kept				✓	Agnostic		Iconsets	
[B]	Privacy dashboards [123, 135, 137, 180, 444]	Privacy	• Description of what data is collected • Description of how data is handled • Purpose statement of data collection				✓	Agnostic		• Timelines • Bar charts • Maps • Network graphs	
		Human agency	• Opportunity to self-assess the system								
		Transparency	• Disclosure of origin and properties of data								
[C]	Risk matrix [4, 232]	Respect for public interest	• Measure of social impact				✓	Agnostic		Two dimensional space (vulnerability vs dependence of the decision)	
		Human Control	• Ability to override the decision made by a system								

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[D]	Moral space [181]	Respect for public interest	• Measure of social impact				✓	Agnostic	Based on human judgement	Three dimensional moral space. Wrongness as a function of intention and harm	
[E]	Social impact assessment [334]	Respect for public interest	• Measure of social impact	✓	✓	✓		Agnostic	Anticipate scenarios		
[F]	Model Tracker interactive visualization [17]	Performance	• Accuracy • False Positive and Negative rates	✓				Classification tasks		• Summary statistics • Confusion matrices • Labels chart • Precision-recall curves • Connector lines to identify similar examples in feature space • Highlighted boxes for correlations between features and target classes	

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[G]	Model cards for models [294]	Perfor-mance	• Accuracy • False Positive and Negative rates • False Discovery and omission rates	✓	✓			Agnostic		• Confidence bars • Bar charts	
		Fairness	• Accuracy across groups • False Positive and Negative rates across groups • False Discovery and omission rates across groups								
[H]	What-if tool [1] [423]	Perfor-mance	• Accuracy • False Positive and Negative Rates • False Discovery and omission rates	✓	✓			Classification tasks, Regression tasks		• Confusion matrices • (Two-dimensional) Histograms • Scatterplots • Summary statistics of datasets • Partial dependence plots	Interactive modules include: list of feature values, inference values, and counterfactual controls.
		Fairness	• Accuracy across groups • False Positive and Negative Rates across groups • False Discovery and omission rates across groups								

[1] https://github.com/pair-code/what-if-tool

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		Transparency	• Disclosure of origin and properties of data								
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[I]	Interactive transfer learning tools [291]	Performance	• Accuracy • False Positive and Negative Rates			✓		Convolutional Neural Networks		• Confusion matrices • Z-scored of each filter • Bar charts • Activation heatmaps • t-SNE clusters	
		Fairness	• Accuracy across groups • False Positive and Negative Rates across groups								
		Transparency	• Disclosure of properties of data								
[J]	Question-Driven XAI Design [256]	Performance	• Accuracy			✓	✓	Agnostic		• Summary statistics (percentage scores) for data explanations and performance metrics • Feature importance • Contrastive explanations	End users were more interested in the limitation of the model: uncertainty
		Fairness	• Accuracy across groups								
		Transparency	• Disclosure of origin and properties of data								

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
		Explainability	• Post-hoc explanations								
[K]	Datasheets for datasets [152]	Transparency	• Description of data generation process • Disclosure of origin properties of models and data	✓	✓			Agnostic		• Summary statistics • Visual examples of datasets (if images, for instance)	
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
		Privacy	• Written declaration of consent • Description of what data is collected • Description of how data is handled • Purpose statement of data collection • Statement of how long the data is kept								
		Fairness	• Election of protected classes								

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		Security	• Membership inference								
[L]	Data centric explanations [21]	Transparency	• Description of data generation process • Disclosure of origin and properties of the models and data			✓	✓	Agnostic		• Interactive list • Q&A format • Pie charts • Bar charts • Process diagrams • timelines • Icons	
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[M]	Example-based explanations [42, 59, 76, 116, 205, 255, 256]	Transparency	• Disclosure of properties of data	✓	✓	✓	✓	Agnostic	• Similar example • Typical example • Counterfactual example	• Example images from dataset if in the visual domain	Normative vs comparative explanations [76]
		Explainability	• Post-hoc explanations								
[N]	Explanation by simplification [42, 205]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Decision rule • Decision tree		

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[O]	Feature relevance explanation [15, 42, 59, 116, 205, 256]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Feature attribute • Feature shape • Feature interaction • Sensitivity / perturbation -based • Saliency maps (visual domain)	• Bar charts • Visualization of element importance, saliency (visual domain)	Usability of saliency maps for non-experts [15]. They should be accompanied by global descriptors
[P]	Contrastive explanations [113, 256, 299]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Example of minimum change that leads to different outcomes		
[Q]	Text-based explanation [42, 53]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• With or without outcome comparison		
[R]	Interactive demonstrations [263]	Explainability	• Post-hoc explanations				✓	Agnostic			

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[S]	Experiential AI [177]	Explainability	• Post-hoc explanations				✓	Agnostic	• Art mediated between computer code and human comprehension		
[T]	Interactive contestations [179, 224]	Contestability	• Mechanisms for users to ask questions and record disagreements with system behavior			✓	✓	Agnostic	• Statements restricted to natural language		
		Human Control	• Ability to override the decision made by the system								
		Human agency	• Opportunity to self-assess the system								
[U]	Challenge justifications provided by operator using the same means [179]	Contestability	• Mechanisms for users to ask questions and record disagreements with system behavior	✓	✓	✓		Agnostic	• Further testing • Verification		

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[V]	Mapping of actors and tasks depending on automation level [77]	Human Control	• Establishment of levels of human discretion during the use of the system	✓	✓	✓	✓	Agnostic		• Relationship diagrams	
[W]	Failure Modes and Effects Analysis [334]	Security	• Threats against integrity (adversarial learning) and mitigation techniques	✓	✓			Agnostic			
		Fairness	• Accuracy across groups • False positives and negatives across groups								
[X]	Aequitas [2] [349]	Fairness	• Accuracy across groups • False Positive and Negative rates across groups • False Discovery and Omission rates across groups • Counterfactual examples	✓	✓			Agnostic			

[2] https://github.com/dssg/aequitas

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[Y]	AI Fairness 360[3] [47]	Performance	• False Positive and Negative rates	✓	✓			Classifiers: logistic regression, random forest classifier and neural networks		• Bar charts • Confidence bars	
		Fairness	• False positive and negative rates across groups • Debiasing algorithms								
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[Z]	Fairlearn[4] [60]	Performance	• Accuracy • False Positive and False Negative rates • Precision and recall rates	✓	✓			Agnostic		• Bar charts • Pie charts	

[3] https://github.com/Trusted-AI/AIF360

[4] https://github.com/fairlearn/fairlearn

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		Fairness	• Accuracy across groups • False negative and false positive rates across groups • Debiasing algorithms								
[AA]	Playbook AI [5] [190]	Human agency	• Give knowledge and tools to comprehend and interact with AI systems • Opportunity to self-assess the system				✓	NLP	Early AI prototyping	• Interactive survey	
[AB]	Counterfit [6]	Security	• Defence against integrity threats • Defence against privacy threats	✓				Agnostic			
[AC]	InterpretML [7] [8] [299, 310]	Explainability	• Interpretability by design • Post-hoc explanations	✓	✓			Both white-box and blackbox models		• Bar charts • Line charts • Decision trees	

[5] https://github.com/microsoft/HAXPlaybook

[6] https://github.com/Azure/counterfit

[7] https://github.com/interpretml/interpret/

[8] https://github.com/interpretml/DiCE

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[AD]	Error analysis dashboard [9]	Non-discrimination	• Analysis of data for potential biases, data quality assessment	✓	✓			Agnostic		• Decision tree • Error heatmap	
		Explainability	• Post-hoc explanations								
		Fairness	• Accuracy across groups								
[AE]	Breakend Impact tracker [10] [178]	Performance	• Estimation of energy consumption • Estimation of GPU memory consumption	✓	✓			Agnostic		• Dot plots • Bar charts	
		Respect for public interest	• Measure of environmental impact								

[9] https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/erroranalysis-dashboard-README.md

[10] https://github.com/Breakend/experiment-impact-tracker

	Means	Value	Manifestation(s)	Stakeholder				Application	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[AF]	Representative contestations [406]	Contestability	• Mechanisms for users to ask questions and record disagreement with system behaviour				✓	Agnostic			

Table 6.3: Mapping of available means for transmitting value-specific manifestations to different stakeholders based on the purpose of their insight and the nature of their knowledge (DT = Development Team; AT = Auditing Team; DE = Data Domain Experts; DS = Decision Subjects). The identification and color code correspond to those on table 6.2. Each means is linked to the value and criteria manifestations that they communicate, the stakeholders that the original papers address, model specificity, deployed approach, visual elements and any additional details.

Appendix B: Chapter 3

Appendix B.1.: Interview Protocol

Thank you very much for taking the time to answer to this interview. My name is ... and I am a researcher at .... As part of my research, I am interested in examining people's perceptions around automated decision-making processes. We will be recording this interview to inform a scientific publication. [Informed consent] Through this interview, in particular, I am interested in knowing more about the way in which people would react to the outcomes automated decision-making processes and whether and how they would like to contest these decisions. We define "contest" as the "act of opposing an action; either because the action is perceived as mistaken or simply wrong". Please feel free to speak without any fear or hesitation. We will anonymize these interviews – meaning your identity – won't be revealed. If you want to say something off the record, please indicate it to me, and I will make sure that this is not included in the analysis. Also, use examples from your experiences or the experiences of your friends and loved ones, which may be relevant to our research. The more elaborate you are, the more we will learn, and this will help us improve these decision-making processes.

PART 1: PARTICIPANTS' BACKGROUND AND EXPERIENCE

Experience with contesting decisions:

1. **First, could tell me a bit about yourself?**
2. **Could you think of any decision made by an organization that you found unfair / treated you unfairly? By any I mean, your most recent experience or the most painful one.**
   (a) If not (they never had a decision made by an organization that they found unfair)
      i. Do you know of any case from someone close to you (e.g., family member, partner, friend)?
   (b) Did you contest this decision? Please, feel free to mention more than one decision if this is the case. Also feel free to describe the commonalities and/or differences from one to another contestation.
   (c) If yes (they decided to contest):
      i. Why did you decide to contest?
      ii. How did you proceed?
   (d) If not (they decided not to contest):
      i. Why did you decide not to contest? What deterred you from raising your voice?

Experience and motivation for renting the house:

1. **Could you tell me about your experience renting your house?**
   (a) Why do you rent your house?
      i. What are the stakes if something goes wrong in your rental?
      ii. Does it have a big impact in your life?
2. **How do you rent your house? Where do you announce the possibility of renting?**

(a) What are the reasons for using these means? What benefits does it bring you?

PART 2: PERCEPTIONS AROUND THE USAGE OF AI

Now let me introduce you the following news:

Show *Prompt1-News.pdf*

1. **How appropriate do you think it is to use Artificial Intelligence systems for detecting housing rental fraud?**
   (a) Can you think of any benefits of using such systems for determining whether a house has been rented illegally?
   (b) (If so) Who is it beneficial for?
   (c) What are the factors / characteristics of an AI system that makes you think this way?
      i. How is this different from human decision-making?

PART 3: GETTING TO KNOW WHAT PARTICIPANTS WOULD LIKE TO CHALLENGE AND HOW

Imagine you received the following letter from the municipality:

Show *Prompt2-Letter.pdf*

Now I would like to capture your reaction to such a scenario. As the letter mentioned, this decision was triggered by an AI system.

1. **How would you react if you received such a letter?**
   (a) How fair do you think this decision-making process is?
   (b) What do you think could be the reasons why you received such a letter?
   (c) How would you solve this problem?
   (d) If you had the chance to contest the warning what aspects of the decision-making process would you like to contest? As mentioned, we define “contest” as the “act of opposing an action; either because the action is perceived as mistaken or simply wrong”.
2. **The letter gives you an op>on to contest this decision by calling the number on the letter, to what extent is this a good contestation process in your opinion?**
   (a) What format would you like to use contest the decision?
   (b) What support (if any) would you seek to help you throughout the process?
   (c) Where would you like to get at the end of the contestation process?
   (d) How would you be satisfied?
3. **How do you think that contestation process allows you to be heard by the municipality?**
   (a) How much of a voice do you think you have in the process? To what extent do you feel empowered?

(b) To what extent would you feel that you have influence over the decision?
(c) To what extent you think that this review process treats you with dignity and respect?

PART 4: ADDRESSING PARTICIPANTS' INSIGHT NEEDS

Now I would like you to look at the following list of details about the system that could be made available to you. I would like you to select those that you **would actually like to know for challenging the decision-making process**.

Show *Prompt3-InformationSheet.pdf*

1. **Are there any of the following aspects of the system and process that you would you like to know more about?**
   (a) What are the reasons for prioritizing this information?
   (b) Why did you prioritize this information A and not B?
2. **How would you use this knowledge in the contestation process?**
   (a) How would you make sense of this information?
   (b) What kind of resources would you use for looking for help?
   (c) Do you think some other people in your social circle would do the same as you intend to do?
3. **Now, how would you now describe a good contestation process?**

PART 5: FINISHING. CLEAN-UP QUESTION

I think this is basically everything that I wanted to ask you about. Do you have anything else you would like to say or any final thoughts?

Do you know of anyone who rents their houses for short periods of time who would be interested in participating in this interview?

Appendix B.2.: Mapping of participants and statements

Our results section avoids naming participants for *each* statement that compose our themes and sub-themes. We, instead, give a sense of the prominence of each statement by using terms such as *a few, many, mostly, generally, unanimously*. We also mention *who* said *what* when a statement is based on comments made by *one* to *three* participants. We decided to do so to improve the readability of our manuscript. Table 6.4 offers a detailed mapping of the participants whose responses led to the statements in our results section. Additionally, we release our codebook, where we include the specific quotes that compose each statement.

Statements	**Participants**
T1.1. Strategizing Information Requests	
Case (1), hypothesized that they had violated the regulation	P1, P3, P4, P5, P8, P10, P11, P12, P13, P15, P16, P17, P18, P19, P20, P21
Case (2), hypothesized that they represented a false positive	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21
Case (2a), hypothesized that they had rented their property out with a license	P3, P6, P8, P19, P10, P13, P15, P18
Case (2b), hypothesized that they had not rented their property out but the system indicated that they did	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P14, P15, P16, P17, P19, P20, P21
Unanimous priority to knowing *why*	P1, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21
Difference between feature-based explanation and decision justification	P1, P4, P6, P7, P8, P9, P10, P12, P15, P16, P17, P19, P20, P21
Municipality should be clear about the reasons behind the decision	P13, P15, P18, P19, P21
Lack of actionable data features	P1, P3, P7, P8, P9, P10, P12, P15, P18
Want to understand the decision basis (policy)	P3, P15, P18
Request for information to know where they stand with respect to law	P1, P5, P15
Want to double-check that the algorithmic decision basis is backed up by relevant policy	P1, P4, P8, P13, P15
Admit that they might not have been aware of the regulation and would accept the first warning if duly motivated	P1, P10, P12, P13, P16
Ask for legal advice	P1, P2, P4, P7, P14, P18, P19
Case (2b), difficult to show proof of innocence	P11
Low AI literacy - uninterested in knowing how the algorithmic system worked	P3, P9, P13, P18, P19
Medium and *High AI literacy* - data-related information regarded as actionable	P1, P2, P4, P5, P6, P7, P8, P10, P11, P15, P16, P20, P21

Statements	**Participants**
Medium AI literacy and *High AI literacy* with no *AI fairness experience*- generally curious to know more about the system	P1, P2, P5, P8, P10, P11, P12, P20, P21
Medium AI literacy and *High AI literacy* with no *AI fairness experience*- generally curious to know more about the system	P1, P2, P5, P8, P10, P11, P12, P20, P21
Medium AI literacy and *High AI literacy* with *no AI fairness experience* - do not know how information about the *how* is useful for contestability	P1, P2, P5, P6, P7, P8, P10, P11, P20, P21
High AI literacy and *experience with AI fairness* - additionally requested information about the model and the development of the system	P4, P15
T1.2. Facilitating Dialogue with Controllers	
Need to turn relevant information into meaningful explanations	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P15, P16, P17, P18, P19
Need to clarify technical jargon	P2, P3, P6, P7, P8, P10, P15, P16
Communication channels should be designed to minimize friction	P1, P2, P3, P4, P5, P6, P8, P9, P10, P12, P13, P14, P15, P17, P18, P20, P21
Effort to understand the information has to be minimal	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P12, P13, P15, P16, P17, P18, P19
Need for information to be relevant for their case, concise, simple and clear	P2, P4, P5, P7, P8, P9, P10, P12, P15, P16, P18
Need to customize information, e.g., to decision subjects' AI literacy	P4, P5, P6, P8, P11, P15, P18, P19, P21
Progressive discovery of information based on relevance	P8, P9, P10, P18
Need for communicative effort through visual explanations or explanatory videos	P8, P16, P19, P21
Interactive explanations	P1, P17, P21
T2.1. Seeking Organizational Support	
Preference towards human reviewer	P2, P5, P6, P12, P13, P16, P17, P19, P21
AI seen as unable to change the output while humans could deal with ill-defined situations	P5, P6, P12, P13, P16, P19
Need for human reviewer to be cooperative, empathetic and proactive	P2, P5, P6, P8, P11, P12, P15, P18, P19, P20, P21
Need for human reviewers to be active listeners	P7, P20
Need for decision subjects to feel understood and heard	P21
Defined the contestation process as a fight	P6, P8, P10
Need for human reviewers to be experts in AI	P2, P3, P4, P6, P7, P8, P11, P12, P15, P16, P20
Power differentials between reviewers and decision subjects	P10, P15, P17
Power differentials accentuated when there is a lack of knowledge on the decision subject's side	P17, P18
Requests for a third party to mediate	P1, P2, P3, P7, P8, P10, P11, P12, P14, P15, P18, P21
Third party could ask questions on the decision subjects' behalf	P15

Statements	**Participants**
Third party could have information about similar cases	P8, P15
Third party should be independent from the municipality, but as informed as developers	P11, P15, P18
Third party should have legal knowledge	P1, P15, P18
Third party should have technical knowledge and help decision subjects move forward	P2, P7, P8, P14, P15
Level of support needed from the third party would depend on the decision subjects' AI literacy	P10, P12, P15, P21
Level of support needed from the third party would depend on the decision subjects' level of satisfaction with the explanation received from the controller	P12
Level of support needed from the third party would depend on the decision subjects' legal knowledge	P10
T2.2. Seeking Peer Support	
General priority to clarify their own individual cases	P3, P5, P6, P7, P8, P9, P10, P11, P14, P15
Case (2b), contesting aspect of the algorithmic system more feasible if done collectively	P11, P15, P18
Similar cases where the algorithmic system made an error - basis for collective contestation	P4, P16, P17
Collective - a place organized by citizens, by people that have gone through this	P18
High AI literacy individuals, or experts could be technical guides	P4, P11, P15
Attracting the attention of the media required	P6, P15
Turning the issues into a political matter required	P6, P15, P18
A collective to help citizens affected by the system to make sense of their situation and act on it	P2, P16
A collective to provide decision subjects insights into similar cases	P4, P7, P14, P16, P17
Collective to enable spotting error patterns across false positives	P3, P4, P7, P11, P13, P16, P17
Collective especially important for people with *low AI literacy* and no immediate social support structures	P2, P18
T3.1. Ensuring Algorithmic Accountability	
Appreciated the right to contest but dealing with errors made by the algorithmic system perceived to be unfair	P9, P10, P15, P20, P21
Burden of showing proof of innocence	P2, P5, P8, P17, P18, P20, P21
Effort needed to make sense of the information that would enable showing proof of innocence	P18
Consensus that correcting AI's mistakes is not the decision subject's responsibility	P4, P7, P13, P15, P18, P21
Requests for *compensations* for the time wasted and effort devoted to contesting	P5, P11
Contesting responsibility of the human controller	P15, P21
Workflow suggested for human controller to contest decisions	P21
Complexity of attributing responsibility if the system is not developed in house	P13

Statements	Participants
Certification as a solution to unburden decision subjects and ensure fair responsibility attribution	P9, P13
T3.2. Fostering Social Transparency	
Requests for transparency about AI development practices within the organizational context of the public administration (*social transparency*)	P4, P15
Requests for participatory development approaches	P1, P11
Importance of probationary periods that do not impact ongoing activities	P1
Nature of public administration makes the choices made during the system development to be the correct ones	P2, P5, P7, P9, P13, P20, P21
Because the public administration is behind the system, assumption that there will be more accountability and diligence	P2, P5, P10, P20, P21
In the public sphere algorithmic decision-making is believed to be more contestable	P2, P21

Table 6.4: Mapping of statements that compose our results section and participants whose remarks led to developing such statements.

Appendix C: Chapter 4

Appendix C.1: Selected quotes

Selected quotes from the preliminary study (S1; see Section 4.4.1) and the main study (S2; see Section 4.4.2). Each quote comes with a reference to the study where the response was collected and to the the participant (P*j*) who gave it.

Q.id	Quote	Participant
Q.1	*"It is unfair for her to be denied based on someone else's previous inability to pay back the loan"*	S1-P42
Q.2	*"Just because some had a similar case as hers, does not prove that she would not be able to pay back the loan."*	S1-P36
Q.3	*"The best explanation gives the largest volume of information including how the decision was made and what amount she could potentially lend"*	S1-P50
Q.4	*"It explains the importance of each factor so she is able to see clearly what factors are most influential"*	S1-P32
Q.5	*"It boils it down to very easy to digest reasons as to why Kim was rejected the loan request"*	S1-P29
Q.6	*"It provides 3 different ways in which Kim could improve her chances of being accepted."*	S1-P33
Q.7	*"She should contest how little impact her employment has on the decisions since this is a big factor"*	S1-P22
Q.8	*"Gender should be contested as is a discriminatory factor. Although all the variables in question are methods for the banks to discriminate against someone, gender is not within a person's control and therefore a bad measure of their character and choices."*	S1-P56
Q.9	*"Artificial intelligence does not take your lifestyle and circumstances into account."*	S1-P46
Q.10	*"It is assessing her by comparing her situation with another with similar salary & credit score & not taking her full circs [circumstances] into consideration."*	S1-P53
Q.11	*"I think there should be a breakdown of what the artificial intelligence looks for and what the decision is based on."*	S2-P5
Q.12	*"They should offer a detailed reason and list of suggested changes she could make to help her in her efforts"*	S2-P218
Q.13	*"It does not tell us enough about how the AI uses the information. The AI is programmed initially by a human. How can I be sure that no bias is involved in this programming of the algorithm? This would be appropriate information to have."*	S2-P8
Q.14	*"If Kim is not familiar with AI then she may not understand the process and view it negatively"*	S2-P135
Q.15	*"[...] each application should be reviewed by a human, not just the ones which have low confidence"*	S2-P179
Q.16	*"Maybe for it to be processed primarily by the AI but secondly by a human before the answer is finalised. This could still be a quick process as the person wouldn't have to spend much time on it but it would mean the decision also had a human input."*	S2-P226

Q.17	*"It is fairer than other options as [it] is quicker than a human decision - [it] allows customers to explore other options"*	S2-P153
Q.18	*"It is fair because with the help of its AI the application process is much faster and efficient"*	S2-P146
Q.19	*"I do think it is fair, it is a quick and easy procedure"*	S2-P182
Q.20	*"It's fair because it can't be biased because it's AI"*	S2-P110
Q.21	*"[...] it may be fair as an algorithm does not take into account factors such as someone's manner or dress which may lead to an unconscious bias for or against an applicant when assessed by a human."*	S2-P8
Q.22	*"It is very fair because all applicants are assessed using the same list of criteria."*	S2-P85
Q.23	*"It takes in essential information needed to evaluate weather a loan is risky from the bank's point of view as a business deal, it doesn't take feelings or emotions, just facts, and applies them to the bank's set criteria with which they are happy to give a loan out to."*	S2-P98
Q.24	*"I think they have asked the correct information to see if an individual could be able to afford to pay back the loan."*	S2-P34
Q.25	*"I think it is fair that it is based on the same factors for everyone but there are circumstances under which more personal information individual to their case should be taken into consideration."*	S2-P51
Q.26	*"The AI system will only deal with data/numbers and won't take into consideration Kim's personal circumstances which could explain why she was rejected in the first place. For example, many lost their jobs due to no fault of their own during the pandemic and fell behind on bills etc. and many have ended up in debt. If this was the case with Kim it wouldn't really be fair based on the circumstances."*	S2-P96
Q.27	*"Everyone is treated the same, but it seems that if a human saw she was only 5% off having the loan, they would have just let it slide."*	S2-P9
Q.28	*"There should be some human to evaluate those cases that are in the obscure region of the cutting-off point."*	S2-P209
Q.29	*"If the person trying to get the loan is rejected within a small margin and appeals I believe they should be able to re-negotiate."*	S2-P185
Q.30	*"They took the human element away, which allows for communication and some compromise."*	S2-P245
Q.31	*"[...] there will always be instances where an AI will get the decision wrong when a person land in a grey area/their circumstances fall into an area where a little compassion is needed."*	S2-P218

Table 6.5: Summary of some of our participants' responses to the open ended questions. S1 = preliminary study, S2 = main study, Pj = index of the participant.

Appendix C.2: Summary of the experimental design

Parameters	Conditions	Descriptions
Explanation	No explanation	*The artificial intelligence system uses some of this information for making the loan decision.*
	With explanations	*In the email received by Kim, an explanation of how the decision-making system has reached the conclusion is included. The email includes the importance that each piece of information provided by Kim had in the final decision. Factors are listed from the most important to the least important factor based on the bank's criteria. The magnitude of the contribution of each piece of information (negative (−) means that it contributed to the rejection decision) is added between brackets:* *Credit Score* (−0.15) > *Loan amount requested* (−0.12)> *Total annual income* (−0.09)> *Loan purpose* (+0.02)> *Employment status* (+0.02)> *Loan amount term (months)* (−0.03)> *Date of birth* (+0.03)> *Co-applicant (if any) income* (+0.01)> *Number of dependents* (−0.07)> *Education* (+0.02) *The email also includes information about scenarios where the individual would have been granted the loan. Kim would have been granted a loan if one of the following scenarios had been true:* • *The loan amount requested had been* 5% *lower* • *The total annual income of the individual had been* 10% *higher* • *The credit score of the individual had been "Very Good"*
Human oversight	No human oversight	*Given the latest technological advances and in an effort to make loan decisions in a timely manner, the loan application process is now fully automated. An artificial intelligence system receives the online requests and evaluates each case. An email is sent to the applicants with the final verdict.*

	With human oversight	*Given the latest technological advances and in an effort to make loan decisions in a timely manner, the loan application process is now hybrid: it combines artificial intelligence with human expertise. This involves a two-step approval process. In the first step, an artificial intelligence system receives the online requests and evaluates each case. If the artificial intelligence system reaches a decision (approve or reject) with a high confidence, an email is sent to the applicant with the final verdict. If the artificial intelligence system has a low confidence over the decision, there is a second step where a human oversees the decision and makes the final verdict and an email is sent to the applicant.*
Contestability	No contestability	*Since the reason for introducing an artificial intelligence system is to handle home loan applications in a timely manner, Kim has no option to request a review of the decision.*
	Contest initial decision	*Kim has decided to appeal the decision and has asked for a review of the process. As part of the review procedure, Kim has the opportunity to make objections about the initial decision and provide any information to support the application. The same artificial intelligence system will then reevaluate the home loan application.*
	Contest decision maker	*Kim has decided to appeal the decision and has asked for a review of the process. As part of the review procedure, Kim has the opportunity to ask for a human to review the process. This human reviewer will make a completely new decision with the information that Kim already provided for the initial decision.*
Task stakes	High stakes	*Buy a house / home loan*
	Low stakes	*Go on holiday / holiday loan*

Table 6.6: Summary of the experimental design.

Appendix D: Chapter 5

Appendix D.1.: Experimental Design

Your city has limited living space; both for citizens and visitors. If a citizen wants to rent out their home on Airbnb to tourists, they need to meet certain requirements. They must also request a license to the municipality. Not everyone adheres to those conditions. The municipality sometimes receives reports that a home has been rented out without meeting the requirements. Until now, a human civil servant would manually investigate the report and find evidence that would help determine whether the reported property was being illegally rented.

Given the shortage of long-term rentals in your city, the municipality has decided to increase its efforts to identify citizens who do not meet the requirements to rent their homes on Airbnb. For this reason the municipality of your city has adopted an Artificial Intelligence system to accelerate the identification of these illegal rentals. With the new system, when a report is filed, the Artificial Intelligence system has access to **[Data provenance]**.

Based on that data, the Artificial Intelligence system **[Model type]** **[Profile]** and it is the first time that this address is reported, a first warning is sent to request the owner to stop renting the property illegally. After this first warning, the owner might face penalties if they fail to adhere to the vacation rental policy.

[We present the diagram of the workflow. We provide an example to illustrate the workflow in practice.]

A few hours ago, a report was filed to complain about a potential case of an illegal holiday rental in 25 Green Hill Street. After retrieving **[Data provenance (...)]**, the evaluation of the Artificial Intelligence is the following:
[Model type] ∩ [Data provenance]

Since **[Model type (...)]**, **[Profile (...)]**. The letter includes a first warning and a request to stop renting the property illegally. It also includes information on how to **[Profile (...2)]** to ask any questions the 25 Green Hill Street owner might have.

Table 6.7: Scenario presented to participants.

Parameters	**Conditions**	**Descriptions**
Profile	Hybrid	*the human civil servant in charge examines the evaluation of the Artificial Intelligence. If, based on the civil servants' judgment, there are clear signs that indicate an illegal holiday rental in this address,* ***(...)*** *the human civil servant in charge has examined the evaluation of the Artificial Intelligence. Based on the civil servant' judgment, there are indeed clear signs that indicate an illegal holiday rental in this address. The human civil servant has, therefore, send a letter to the property owner of 25 Green Hill Street.* ***(...2)*** *contact the human civil servant in charge,*
	Fully-automated	*-* ***(...)*** *the evaluation of the Artificial Intelligence system has led to a letter to be sent to the property owner of 25 Green Hill Street.* ***(...2)*** *interact with the Artificial Intelligence system*
Model type	Probabilistic	*calculates the probability of a property being illegally rented on the reported address. If the probability is high,* ***(...)*** *the probability of this property being illegally rented is high,*
	Rule-based	*evaluates through a rule-based system whether the reported address meets the conditions of illegal holiday rental. If relevant conditions are met that indicate an illegal holiday rental in this property,* ***(...)*** *relevant conditions are met that indicate an illegal holiday rental in this property,*
Data provenance	Publicly available databases	*the public registry, where it retrieves information about prior illegal housing cases, about the building and about the identity and housing rights of the residents.* ***(...)*** *information from the public registry*
	Non publicly available data sources	*the camera footage of the doorbell in the building. If the doorbell has no camera, then it accesses the footage of the nearest street camera. Thanks to this footage, the AI identifies the flow of people accessing the building.* ***(...)*** *footage from the cameras*
Model ∩ Data	Probabilistic ∩ Public	*"The property in 25 Green Hill Street has a high probability probability of being an illegal holiday rental. According to the information in the public registry, the following factors determine the high probability:* • *Street code +++* • *Anonymous reporter +++* • *Number of rooms ++* • *Date of residence in the address +"* *(+) means that this factors contributed to getting a high probability. The more (+) signs, the bigger the impact of that factor on getting a high probability.*

Probabilistic Non-public	∩	*"The property in 25 Green Hill Street has a high probability probability of being an illegal holiday rental. According to the information obtained from the camera in the last month, the following factors determine the high probability:* • *Total number of suitcases detected entering the building +++* • *Total number of non-regular residents entering the building +++* • *Flow of people during weekends and holidays ++* • *Frequency of access of people during working hours +"* *(+) means that this factors contributed to getting a high probability. The more (+) signs, the bigger the impact of that factor on getting a high probability.*
Rule-based Public	∩	*"The property in 25 Green Hill Street meets the conditions for being flagged as an illegal holiday rental. According to the information in the public registry, the following conditions were met:* • *The property is located in a highly touristic area of the city* • *The complaint is not anonymous, it comes from the neighbour nextdoor* • *The property has more than 2 rooms* • *The property owner is not registered in this address and has several other properties"*
Rule-based Non-public	∩	*"The property in 25 Green Hill Street meets the conditions for being flagged as an illegal holiday rental. According to the information obtained from the camera, the following conditions were met in the last month:* • *Total number of suitcases detected entering the building > 15* • *Total number of non-regular residents entering the building > 50* • *Flow of people during weekends and holidays > 5 people entering the building on average every 30 minutes during the day* • *Flow of people during working hours > 3 people entering the building on average every hour"* *These conditions apply to this particular building based on its size and factors such as the presence of other Airbnb-s in the building.*

Table 6.8: Experimental design. Depending on the scenario assigned to each participant (e.g., profile = *hybrid*, model type = *probabilistic*, data provenance = *publicly available data sources*), the descriptions corresponding to those conditions were introduced in the scenario presented in table 6.7.

Appendix D.2.: Measurement Tools

Measurements

A. Items to measure *perceived ability.* Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. **[Decision-maker]**[11] has the competence to include all necessary information for making decisions about illegal holiday rentals.
2. **[Decision-maker]** is able to process all data necessary for making decisions about illegal holiday rentals.
3. **[Decision-maker]** is able to consider all necessary data when making decisions about illegal holiday rentals.
4. **[Decision-maker]** is capable of flexibly considering different circumstances when making decisions about illegal holiday rentals.
5. **[Decision-maker]** has the competence to adapt its decision to different circumstances.
6. **[Decision-maker]** is able to react flexibly to circumstances in the decision-making process.

B. Items to measure *perceived benevolence.* Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. **[Decision-maker]** will take care of the welfare of the owner of 25 Green Hill Street.
2. **[Decision-maker]** will consider the needs and desires of the owner of 25 Green Hill Street.
3. **[Decision-maker]** will act on the best interest if the owner of 25 Green Hill Street.
4. **[Decision-maker]** will look out what is important for the owner of 25 Green Hill Street.
5. **[Decision-maker]** will go out of its way to help the owner of 25 Green Hill Street.

C. Items to measure *perceived integrity.* Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. **[Decision-maker]** acts with a strong sense of justice.
2. **[Decision-maker]** acts in an honest way.
3. **[Decision-maker]** is fair when identifying illegal holiday rentals.
4. The behaviours and decisions coming out of **[Decision-maker]** are not very consistent (r).
5. I like the values and purposes behind having a **[Decision-maker]** for identifying illegal holiday rentals.
6. Sound principles guide the behaviour of **[Decision-maker]**.

D. Item to measure *perceived fairness.* Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. Overall the decision-making process for identifying illegal holiday rentals set up by the municipality is fair.

Descriptive and control variables

A. Questionnaire for determining *age range.*

What is your age range?

- A1: 0-18
- A2: 19-25
- A3: 26-35
- A4: 36-50

[11] **[Decision-maker]** is either *"The Artificial Intelligence system"* or *"The human civil servant (by) using the Artificial Intelligence system and their own judgment"* depending on the condition that each participant gets.

- A5: 50-80
- A6: 80+

B. Questionnaire for determining *level of education*.

What is the highest level of school that you have completed or the highest degree you have received?

- A1: High school incomplete or less.
- A2: High school graduate or GED (includes technical / vocational training that does not award college credit)
- A3: Some college (some community college, associate's degree).
- A4: Four year college degree / bachelor's degree
- A5: Some postgraduate or professional schooling, no postgraduate degree
- A6: Postgraduate or professional degree, including master's, doctorate, medical or law degree

C. Items to determine experience as *lessee of short-term rentals*. Assessed as a yes/no question.

1. I have rented my house out for short-term rentals (for example, Airbnb) and I had a license for it.
2. I have rented my house out for short-term rentals and I did not have a license for it.

D. Items to measure *AI literacy*. Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. I have a good knowledge in the field of *artificial intelligence*.
2. My current employment includes working with *artificial intelligence*.
3. I am confident interacting with *artificial intelligence*.
4. I understand what the term *artificial intelligence* means.

E. Items to measure *affinity to technology*. Assessed on a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. I like to occupy myself in greater details with technical systems (systems that include some technology: computing systems, electronic gadgets, mechanisms)
2. I like testing functions of new technical systems.
3. It is enough for me that a technical system works; I don't care about how or why (r)[12].
4. It is enough for me to know the basic functions of a technical system (r).

F. Items to measure *personal experience with short-term rentals*. Assessed in a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. I am aware of human civil servants identifying illegal holiday rentals.
2. I am aware of artificial intelligence systems detecting illegal holiday rentals.

G. Items to measure *personal experience with public administration*. Assessed in a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. I have a good experience dealing with the human civil servants in the public administration.

H. Item to measure *affinity to short-term rental policy*. Assessed in a seven-point Likert scale (1 = completely disagree, 7 = completely agree).

1. It is acceptable that the municipality enforces a policy to identify and penalize short-term rentals like Airbnb(s) that are not officially registered.

[12] Reverse coded

I. Item to measure *perceived task complexity.* Assessed in a seven-point Likert scale (1 = very low in complexity, 7 = very high in complexity).

1. How complex do you think it is to identify illegal holiday rentals?

J. Open-ended questions.

1. Do you think [**Decision-maker**] is capable of correctly identifying illegal holiday rentals? Why?
2. Do you think the [**Decision-maker**] will try to help the 25 Green Hill Street owner? Why?
3. Do you think it is right that the municipality relies on [**Decision-maker**] for the decision-making process? Why?

Bibliography

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. “Roles for computing in social change”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2020, pp. 252–260. ISBN: 9781450369367. DOI: 10.1145/3351095.3372871.

[2] J. Stacy Adams. “Inequity in social exchange”. In: *Advances in experimental social psychology*. Vol. 2. Academic Elsevier, 1965, pp. 267–299.

[3] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. “Auditing black-box models for indirect influence”. In: *Knowledge and Information Systems* 54.1 (2018), pp. 95–122. ISSN: 0219-3116. DOI: 10.1007/s10115-017-1116-3. URL: https://doi.org/10.1007/s10115-017-1116-3.

[4] AI Ethics Impact Group (AIEIG). *From Principles to Practice An interdisciplinary framework to operationalise AI ethics*. 2020. URL: https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf.

[5] Evgeni Aizenberg and Jeroen van den Hoven. “Designing for human rights in AI”. In: *Big Data & Society* 7.2 (July 2020), p. 2053951720949566. ISSN: 2053-9517. DOI: 10.1177/2053951720949566. URL: https://doi.org/10.1177/2053951720949566.

[6] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. “Elessar: Ethics in Norm-Aware Agents”. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS ’20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2020, pp. 16–24. ISBN: 9781450375184.

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

[8] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. “Tensions in transparent urban AI: designing a smart electric vehicle charge point”. In: *AI & SOCIETY* (Mar. 2022). ISSN: 0951-5666. DOI: 10.1007/s00146-022-01436-9.

[9] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. “Contestable AI by Design: Towards a Framework”. In: *Minds and Machines* (Aug. 2022). ISSN: 0924-6495. DOI: 10.1007/s11023-022-09611-z.

[10] Kars Alfrink, T. Turel, A. I. Keller, N. Doorn, and G. W. Kortuem. "Contestable City Algorithms". In: International Conference on Machine Learning Workshop, July 2020.

[11] Saja Aljuneidi, Wilko Heuten, Larbi Abdenebaoui, Maria K Wolters, and Susanne Boll. "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*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642535. URL: https://doi.org/10.1145/3613904.3642535.

[12] Saja Aljuneidi, Wilko Heuten, Markus Tepe, and Susanne Boll. "Did that AI just Charge me a Fine? Citizens' Perceptions of AI-based Discretion in Public Administration". In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. New York, NY, USA: ACM, Sept. 2023, pp. 57–67. ISBN: 9798400701160. DOI: 10.1145/3582515.3609518.

[13] Ali Alkhatib and Michael Bernstein. "Street-Level Algorithms". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300760.

[14] Marco Almada. "Human intervention in automated decision-making". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. New York, NY, USA: ACM, June 2019, pp. 2–11. ISBN: 9781450367547. DOI: 10.1145/3322640.3326699.

[15] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. "Evaluating saliency map explanations for convolutional neural networks". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, Mar. 2020. ISBN: 9781450371186. DOI: 10.1145/3377325.3377519.

[16] Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. "The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020". In: *EPJ Data Science* 10.1 (2021), p. 15. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-021-00271-0. URL: https://doi.org/10.1140/epjds/s13688-021-00271-0.

[17] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. "ModelTracker". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2015. ISBN: 9781450331456. DOI: 10.1145/2702123.2702509.

[18] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 289–295. ISBN: 9781450363242. DOI: 10.1145/3306618.3314243. URL: https://doi.org/10.1145/3306618.3314243.

[19] Amnesty International. *Xenofobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal.* 2021. URL: https://www.amnesty.nl/content/uploads/2021/10/20211014_FINAL_Xenophobic-Machines.pdf?x25337..

[20] Access Now Amnesty International. *Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems.* 2018. URL: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

[21] Ariful Islam Anik and Andrea Bunt. "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.* New York, NY, USA: ACM, May 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445736.

[22] Karl Aquino. "Relationships among pay inequity, perceptions of procedural justice, and organizational citizenship". In: *Employee Responsibilities and Rights Journal* 8.1 (Mar. 1995), pp. 21–33. ISSN: 0892-7545. DOI: 10.1007/BF02621253.

[23] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. "In AI we trust? Perceptions about automated decision-making by artificial intelligence". In: *AI & SOCIETY* 35.3 (Sept. 2020), pp. 611–623. ISSN: 0951-5666. DOI: 10.1007/s00146-019-00931-w.

[24] Sherry R Arnstein. "A Ladder of Citizen Participation". In: *Journal of the American Planning Association* 85.1 (Jan. 2019), pp. 24–34. ISSN: 0194-4363. DOI: 10.1080/01944363.2018.1559388. URL: https://doi.org/10.1080/01944363.2018.1559388.

[25] Alejandro Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[26] Mahmoud Assran, Joshua Romoff, Nicolas Ballas, Joelle Pineau, and Michael Rabbat. "Gossip-based Actor-Learner Architectures for Deep Reinforcement Learning". In: (June 2019).

[27] Shubham Atreja, Jane Im, Paul Resnick, and Libby Hemphill. "AppealMod: Shifting Effort from Moderators to Users Making Appeals". In: (Jan. 2023).

[28] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine experiment". In: *Nature* 563.7729 (Nov. 2018), pp. 59–64. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0637-6.

[29] Simone Bae, Reeva Lederman, and Tingru Cui. "Understanding User Perception of Explainable Algorithmic Decision-Making Systems: A Systematic Literature Review". In: (2022).

[30] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. "Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems". In: (Mar. 2021).

[31] Agathe Balayn and Seda Gürses. *Beyond Debiasing: Regulating AI and its inequalities*. 2021. URL: https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/.

[32] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. "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". In: *The VLDB Journal* 30.5 (Sept. 2021), pp. 739–768. ISSN: 1066-8888. DOI: 10.1007/s00778-021-00671-8.

[33] Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, Mireia Yurrita, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, et al. "From Stem to Stern: Contestability Along AI Value Chains". In: *Workshop at the 27th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. (2024).

[34] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. "How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?" In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–16. ISBN: 9781450391573. DOI: 10.1145/3491102.3517474.

[35] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. "An Empirical Exploration of Trust Dynamics in LLM Supply Chains". In: *Proceedings of the Trust and Reliance in Evolving Human-AI Workflows Workshop at CHI Conference on Human Factors in Computing Systems*. (2024).

[36] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. ""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*. New York, NY, USA: ACM, Aug. 2023, pp. 482–495. ISBN: 9798400702310. DOI: 10.1145/3600211.3604674.

[37] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. "Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 116–128. ISBN: 9781450383097. DOI: 10.1145/3442188.3445875. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3442188.3445875.

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

[39] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 2429–2437. ISSN: 2374-3468. DOI: 10.1609 /aaai.v33i01.33012429.

[40] Jeffrey Bardzell, Shaowen Bardzell, and Erik Stolterman. "Reading critical designs". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2014, pp. 1951–1960. ISBN: 9781450324731. DOI: 10.1145/2556288.2557137.

[41] Julian Barling and Michelle Phillips. "Interactional, Formal, and Distributive Justice in the Workplace: An Exploratory Study". In: *The Journal of Psychology* 127.6 (Nov. 1993), pp. 649–656. ISSN: 0022-3980. DOI: 10.1080/00223980.1993.9914904.

[42] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/J.INFFUS.2019.12.012.

[43] Christine Barter and Emma Renold. "'I wanna tell you a story': Exploring the application of vignettes in qualitative research with children and young people". In: *International Journal of Social Research Methodology* 3.4 (Jan. 2000), pp. 307–323. ISSN: 1364-5579. DOI: 10.1080/13645570050178594.

[44] Patrick Bedué and Albrecht Fritzsche. "Can we trust AI? an empirical investigation of trust requirements and guide to successful AI adoption". In: *Journal of Enterprise Information Management* 35.2 (2022), pp. 530–549.

[45] Gianluca Bei and Filippo Celata. "Challenges and effects of short-term rentals regulation". In: *Annals of Tourism Research* 101 (July 2023), p. 103605. ISSN: 01607383. DOI: 10.1016/j.annals.2023.103605.

[46] Uladzislau Belavusau and Kristin Henrard. "A Bird's Eye View on EU Anti-Discrimination Law: The Impact of the 2000 Equality Directives". In: *German Law Journal* 20.05 (July 2019), pp. 614–636. ISSN: 2071-8322. DOI: 10.1017/glj.2019.53.

[47] R K E Bellamy, K Dey, M Hind, S C Hoffman, S Houde, K Kannan, P Lohia, J Martino, S Mehta, A Mojsilović, S Nagar, K Natesan Ramamurthy, J Richards, D Saha, P Sattigeri, M Singh, K R Varshney, and Y Zhang. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 1–4. DOI: 10.1147/JRD.2019.2942287.

[48] Izak Benbasat and Weiquan Wang. "Trust In and Adoption of Online Recommendation Agents". In: *Journal of the Association for Information Systems* 6.3 (Mar. 2005), pp. 72–101. ISSN: 15369323. DOI: 10.17705/1jais.00065.

[49] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3442188.3445922.

[50] Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018). ISSN: 2307-387X. DOI: 10.1162/tacl{_}a{_}00041.

[51] Roni Berger. "Now I see it, now I don't: researcher's position and reflexivity in qualitative research". In: *Qualitative Research* 15.2 (Apr. 2015), pp. 219–234. ISSN: 1468-7941. DOI: 10.1177/1468794112468475.

[52] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art". In: (Mar. 2017).

[53] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. "Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–13. ISBN: 9781450380966. DOI: 10.1145/3411764.3445365.

[54] Niels van Berkel, Eleftherios Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. "A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond". In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. New York, NY, USA: ACM, Oct. 2020, pp. 1–12. ISBN: 9781450375795. DOI: 10.1145/3419249.3420106.

[55] Niels van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. "The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT". In: *International Journal of Human-Computer Studies* 170 (Feb. 2023), p. 102954. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2022.102954.

[56] R.J. Bies and J. F. Moag. "Interactional Justice: Communication Criteria of Fairness. " In: *Research on Negotiations in Organizations* 1 (1986), pp. 43–55.

[57] Robert J. Bies and Debra L. Shapiro. "Interactional fairness judgments: The influence of causal accounts". In: *Social Justice Research* 1.2 (June 1987), pp. 199–218. ISSN: 0885-7466. DOI: 10.1007/BF01048016.

[58] Battista Biggio and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". In: *Pattern Recognition* 84 (Dec. 2018), pp. 317–331. ISSN: 00313203. DOI: 10.1016/j.patcog.2018.07.023.

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

[60] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft, May 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[61] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. "The Values Encoded in Machine Learning Research". In: (June 2021).

[62] Alice Namuli Blazevic, Patrick Mugalula, and Andrew Wandera. "Towards Operationalizing the Data Protection and Privacy Act 2020: Understanding the Draft Data Protection and Privacy Regulations, 2020". In: *SSRN Electronic Journal* (2021). ISSN: 1556-5068. DOI: 10.2139/ssrn.3776353.

[63] Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: (May 2020).

[64] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. "SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 150–159. ISBN: 9781450361255. DOI: 10.1145/3287560.3287583. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3287560.3287583.

[65] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. Sage, 1998.

[66] C. Malik Boykin, Sophia T. Dasch, Vincent Rice Jr., Venkat R. Lakshminarayanan, Taiwo A. Togun, and Sarah M. Brown. "Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: ACM, Oct. 2021, pp. 1–9. ISBN: 9781450385534. DOI: 10.1145/3465416.3483302.

[67] Virginia Braun and Victoria Clarke. "Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches". In: *Counselling and Psychotherapy Research* 21.1 (Mar. 2021), pp. 37–47. ISSN: 1473-3145. DOI: 10.1002/capr.12360.

[68] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (Jan. 2006), pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa.

[69] Sarah Brayne. "Big data surveillance: The case of policing". In: *American sociological review* 82.5 (2017), pp. 977–1008.

[70] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. "Toward Algorithmic Accountability in Public Services". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300271.

[71] Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant. “Of, for, and by the people: the legal lacuna of synthetic persons”. In: *Artificial Intelligence and Law* 25.3 (Sept. 2017), pp. 273–291. ISSN: 0924-8463. DOI: 10.1007/s10506-017-9214-9.

[72] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021), pp. 1–21. ISSN: 2573-0142. DOI: 10.1145/3449287.

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

[74] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Sept. 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[75] Jenna Burrell. “How the machine ‘thinks’: Understanding opacity in machine learning algorithms.” In: *Big data & society* 3.1 (2016).

[76] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. “The effects of example-based explanations in a machine learning interface”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, Mar. 2019. ISBN: 9781450362726. DOI: 10.1145/3301275.3302289.

[77] Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci, and Bart Van Arem. “A human centric framework for the analysis of automated driving systems based on meaningful human control”. In: *Theoretical Issues in Ergonomics Science* 21.4 (2019), pp. 478–506. DOI: 10.1080/1463922X.2019.1697390. URL: https://www.tandfonline.com/action/journalInformation?journalCode=ttie20.

[78] Scott Allen Cambo and Darren Gergle. “Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science”. In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–19. ISBN: 9781450391573. DOI: 10.1145/3491102.3501998.

[79] Arturo Casadevall and Ferric C. Fang. “Reforming Science: Methodological and Cultural Reforms”. In: *Infection and Immunity* 80 (3 Mar. 2012), pp. 891–896. ISSN: 0019-9567. DOI: 10.1128/IAI.06183-11.

[80] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. “Task-Dependent Algorithm Aversion”. In: *Journal of Marketing Research* 56.5 (Oct. 2019), pp. 809–825. ISSN: 0022-2437. DOI: 10.1177/0022243719851788.

[81] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van der Hoven, Deborah Forster, and Reginald L. Lagendijk. “Meaningful human control: actionable properties for AI system development”. In: *AI Ethics* 3 (2023), pp. 241–255. URL: https://doi.org/10.1007/s43681-022-00167-3.

[82] David Chan. “Perceptions of fairness”. In: *Research Collection School of Social Sciences* (2011).

[83] Kyla Chasalow and Karen Levy. “Representativeness in Statistics, Politics, and Machine Learning”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 77–89. ISBN: 9781450383097. DOI: 10.1145/3442188.3445872. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3442188.3445872.

[84] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. “Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300789. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3290605.3300789.

[85] China Electronics Standardization Institute. *Original CSET Translation of "Artificial Intelligence Standardization White Paper"*. Jan. 2018. URL: https://cset.georgetown.edu/research/artificial-intelligence-standardization-white-paper/.

[86] Alexandra Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: (Oct. 2016).

[87] C W Churchman. “Free for All”. In: *Management Science* 14.4 (Dec. 1967), B-141-B–146. ISSN: 0025-1909. DOI: 10.1287/mnsc.14.4.B141.

[88] Nazli Cila. “Designing Human-Agent Collaborations: Commitment, responsiveness, and support”. In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–18. ISBN: 9781450391573. DOI: 10.1145/3491102.3517500.

[89] Nazli Cila, Gabriele Ferri, Martijn de Waal, Inte Gloerich, and Tara Karpinski. “The Blockchain and the Commons: Dilemmas in the Design of Local Platforms”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376660.

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

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

[92] Jennifer Cobbe, Michael Veale, and Jatinder Singh. “Understanding accountability in algorithmic supply chains”. In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2023, pp. 1186–1197. ISBN: 9798400701924. DOI: 10.1145/3593013.3594073.

[93] Robert Collins and Johan Redström. “The Contestation Café”. In: *Nordic Human-Computer Interaction Conference*. New York, NY, USA: ACM, Oct. 2022, pp. 1–1. ISBN: 9781450396998. DOI: 10.1145/3546155.3547290.

[94] Jason A Colquitt and Jessica B Rodell. “Measuring Justice and Fairness”. In: *The Oxford Handbook of Justice in the Workplace*. Oxford University Press, Sept. 2015. ISBN: 9780199981410. DOI: 10.1093/oxfordhb/9780199981410.013.0008. URL: https://doi.org/10.1093/oxfordhb/9780199981410.013.0008.

[95] Jason A. Colquitt. “On the dimensionality of organizational justice: A construct validation of a measure.” In: *Journal of Applied Psychology* 86.3 (June 2001), pp. 386–400. ISSN: 1939-1854. DOI: 10.1037/0021-9010.86.3.386.

[96] Jason A. Colquitt and Jessica B. Rodell. “Justice, Trust, and Trustworthiness: A Longitudinal Analysis Integrating Three Theoretical Perspectives”. In: *Academy of Management Journal* 54.6 (Dec. 2011), pp. 1183–1206. ISSN: 0001-4273. DOI: 10.5465/amj.2007.0572.

[97] JOHN COOK and TOBY WALL. “New work attitude measures of trust, organizational commitment and personal need non-fulfilment”. In: *Journal of Occupational Psychology* 53.1 (Mar. 1980), pp. 39–52. ISSN: 0305-8107. DOI: 10.1111/j.2044-8325.1980.tb00005.x.

[98] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[99] Kate Crawford and Trevor Paglen. *Excavating AI: The Politics of Training Sets for Machine Learning*. 2019.

[100] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. “Casting out Demons: Sanitizing Training Data for Anomaly Sensors”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, May 2008, pp. 81–95. ISBN: 978-0-7695-3168-7. DOI: 10.1109/SP.2008.11.

[101] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. “Interactive Model Cards: A Human-Centered Approach to Model Documentation”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 427–439. ISBN: 9781450393522. DOI: 10.1145/3531146.3533108.

[102] Russell Cropanzano. *Justice in the Workplace: From theory To Practice*. Vol. 2. 2012.

[103] Alexander D’Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. “Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 525–534. ISBN: 9781450369367. DOI: 10.1145/3351095.3372878. URL: https://doi.org/10.1145/3351095.3372878.

[104] Steven Dalton, Iuri Frosio, and Michael Garland. “Accelerating Reinforcement Learning through GPU Atari Emulation”. In: (July 2019).

[105] Dasha Simons. *Design for fairness in AI: Cooking a fair AI Dish*. Tech. rep. Delft University of Technology. Graduation project. MSc in Strategic Product Design., 2019. URL: http://resolver.tudelft.nl/uuid:5a116c17-ce0a-4236-b283-da6b8545628c.

[106] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets". In: *Proceedings of the Third Workshop on Abusive Language Online*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. DOI: 10.18653/v1/W19-3504.

[107] Janet Davis and Lisa P. Nathan. "Value Sensitive Design: Applications, Adaptations, and Critiques". In: *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer Netherlands, 2015, pp. 11–40. DOI: 10.1007/978-94-007-6970-0{_}3.

[108] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. "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*. New York, NY, USA: ACM, Apr. 2023, pp. 1–18. ISBN: 9781450394215. DOI: 10.1145/3544548.3581026.

[109] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. "Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 473–484. ISBN: 9781450393522. DOI: 10.1145/3531146.3533113.

[110] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets". In: (July 2020). URL: https://arxiv.org/abs/2007.07399.

[111] Morton Deutsch. "Equity, equality, and need: What determines which value will be used as the basis of distributive justice?" In: *Journal of Social Issues* (1975), pp. 137–149.

[112] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. "Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior". In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–19. ISBN: 9781450391573. DOI: 10.1145/3491102.3517441.

[113] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: (Feb. 2018).

[114] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), pp. 114–126. ISSN: 1939-2222. DOI: 10.1037/xge0000033.

[115] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. "Measuring and Mitigating Unintended Bias in Text Classification". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, Dec. 2018, pp. 67–73. ISBN: 9781450360128. DOI: 10.1145/3278721.3278729.

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

[117] Ravit Dotan and Smitha Milli. “Value-laden Disciplinary Shifts in Machine Learning”. In: (Dec. 2019).

[118] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Oct. 2021), pp. 48–59. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/18939.

[119] Tim Draws, Zoltán Szlávik, Benjamin Timmermans, Nava Tintarev, Kush R. Varshney, and Michael Hind. “Disparate Impact Diminishes Consumer Trust Even for Advantaged Users”. In: (Jan. 2021). DOI: 10.1007/978-3-030-79460-6{_}11. URL: http://arxiv.org/abs/2101.12715%20http://dx.doi.org/10.1007/978-3-030-79460-6_11.

[120] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. “Fairness Through Awareness”. In: (Apr. 2011).

[121] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. “Fairness through Awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS ’12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2090236.2090255.

[122] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: 2006, pp. 265–284. DOI: 10.1007/11681878{_}14.

[123] Julia Earp and Jessica Staddon. “"I had no idea this was a thing"”. In: *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*. New York, NY, USA: ACM, Dec. 2016, pp. 79–86. ISBN: 9781450348263. DOI: 10.1145/3046055.3046062.

[124] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. “FaiRecSys: mitigating algorithmic bias in recommender systems”. In: *International Journal of Data Science and Analytics* 9.2 (2020), pp. 197–213. ISSN: 2364-4168. DOI: 10.1007/s41060-019-00181-5. URL: https://doi.org/10.1007/s41060-019-00181-5.

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

[126] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. “Expanding Explainability: Towards Social Transparency in AI systems”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–19. ISBN: 9781450380966. DOI: 10.1145/3411764.3445188.

[127] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. "The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations". In: (July 2021).

[128] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. "The Algorithmic Imprint". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1305–1317. ISBN: 9781450393522. DOI: 10.1145/3531146.3533186. URL: https://doi.org/10.1145/3531146.3533186.

[129] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. "Operationalizing Human-Centered Perspectives in Explainable AI". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380959.

[130] Carsten Eickhoff. "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, Feb. 2018, pp. 162–170. ISBN: 9781450355810. DOI: 10.1145/3159652.3159654.

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

[132] European Commission. *Ethics guidelines for trustworthy AI*. 2019. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[133] European Commission. *Proposal for 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*. 2021.

[134] European Commission. *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*. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[135] Florian M. Farke, David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. "Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity (Extended Version)". In: (May 2021).

[136] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences." In: *Behavior research methods* 39.2 (May 2007), pp. 175–91. ISSN: 1554-351X. DOI: 10.3758/bf03193146.

[137] Simone Fischer-Hübner, Julio Angulo, Farzaneh Karegar, and Tobias Pulls. "Transparency, Privacy and Trust – Technology for Tracking and Controlling My Data Disclosures: Does This Work?" In: 2016, pp. 3–14. DOI: 10.1007/978-3-319-41354-9{_}1.

[138] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI”. In: *SSRN Electronic Journal* (2020). ISSN: 1556-5068. DOI: 10.2139/ssrn.3518482.

[139] Luciano Floridi. “Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki”. In: *Philosophy & Technology* 33.4 (Dec. 2020), pp. 541–546. ISSN: 2210-5433. DOI: 10.1007/s13347-020-00434-3.

[140] Luciano Floridi. “Establishing the rules for building trustworthy AI”. In: *Nature Machine Intelligence* 1.6 (May 2019), pp. 261–262. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0055-y.

[141] Luciano Floridi. “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical”. In: *Philosophy & Technology* 32.2 (June 2019). ISSN: 2210-5433. DOI: 10.1007/s13347-019-00354-x.

[142] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. In: *Minds and Machines* 28.4 (Dec. 2018). ISSN: 0924-6495. DOI: 10.1007/s11023-018-9482-5.

[143] Thomas Franke, Christiane Attig, and Daniel Wessel. “A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale”. In: *International Journal of Human–Computer Interaction* 35.6 (Apr. 2019), pp. 456–467. ISSN: 1044-7318. DOI: 10.1080/10447318.2018.1456150.

[144] Christopher Frauenberger. “Critical Realist HCI”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2016, pp. 341–351. ISBN: 9781450340823. DOI: 10.1145/2851581.2892569.

[145] Christopher Frauenberger, Marjo Rauhala, and Geraldine Fitzpatrick. “In-Action Ethics: Table 1.” In: *Interacting with Computers* (June 2016). ISSN: 0953-5438. DOI: 10.1093/iwc/iww024.

[146] W. Fred van Raaij and Theo M.M. Verhallen. “Domain-specific Market Segmentation”. In: *European Journal of Marketing* 28.10 (Oct. 1994), pp. 49–66. ISSN: 0309-0566. DOI: 10.1108/03090569410075786.

[147] Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods”. In: *Foundations and Trends® in Human–Computer Interaction* 11.2 (2017), pp. 63–125. ISSN: 1551-3955. DOI: 10.1561/1100000015.

[148] Elena Fumagalli, Sarah Rezaei, and Anna Salomons. “OK computer: Worker perceptions of algorithmic recruitment”. In: *Research Policy* 51.2 (Mar. 2022), p. 104420. ISSN: 00487333. DOI: 10.1016/j.respol.2021.104420.

[149] Georges Gaillard. “La conflictualité : une modalité de lien où s’arrime la destructivité humaine”. In: *Connexions* 106.2 (2016), p. 71. ISSN: 0337-3126. DOI: 10.3917/cnx.106.0071.

[150] Yanjie Gao, Yu Liu, Hongyu Zhang, Zhengxian Li, Yonghao Zhu, Haoxiang Lin, and Mao Yang. “Estimating GPU memory consumption of deep learning models”. In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York, NY, USA: ACM, Nov. 2020, pp. 1342–1352. ISBN: 9781450370431. DOI: 10.1145/3368089.3417050.

[151] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. “Estimation of energy consumption in machine learning”. In: *Journal of Parallel and Distributed Computing* 134 (Dec. 2019), pp. 75–88. ISSN: 07437315. DOI: 10.1016/j.jpdc.2019.07.007.

[152] Timnit Gebru, Google Jamie Morgenstern, Briana Vecchione, and Jennifer Wortman Vaughan. “Datasheets for Datasets”. In: 2020.

[153] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. “Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 325–336. ISBN: 9781450369367. DOI: 10.1145/3351095.3372862. URL: https://doi.org/10.1145/3351095.3372862.

[154] Meric Altug Gemalmaz and Ming Yin. “Understanding Decision Subjects’ Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, July 2022, pp. 295–306. ISBN: 9781450392471. DOI: 10.1145/3514094.3534201.

[155] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. “Measuring Social Biases of Crowd Workers using Counterfactual Queries”. In: (Apr. 2020).

[156] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. “The false hope of current approaches to explainable artificial intelligence in health care”. In: *The Lancet Digital Health* 3.11 (Nov. 2021), e745–e750. ISSN: 25897500. DOI: 10.1016/S2589-7500(21)00208-9.

[157] Amir Globerson and Sam Roweis. “Nightmare at test time”. In: *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. New York, New York, USA: ACM Press, 2006, pp. 353–360. ISBN: 1595933832. DOI: 10.1145/1143844.1143889.

[158] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: (Dec. 2014).

[159] Google. *AI at Google: Our Principles*. 2018. URL: https://www.blog.google/technology/ai/ai-principles/.

[160] Ben Green. “The flaws of policies requiring human oversight of government algorithms”. In: *Computer Law & Security Review* 45 (July 2022), p. 105681. ISSN: 02673649. DOI: 10.1016/j.clsr.2022.105681.

[161] Ben Green and Lily Hu. “The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning”. In: *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*. Stockholm, Sweden, 2018.

[162] J. Greenberg. “The social side of fairness: Interpersonal and informational classes of organizational justice.” In: *Justice in the workplace: Approaching fairness in human resource management.* (1993), pp. 79–103.

[163] Jerald Greenberg. “A Taxonomy of Organizational Justice Theories”. In: *The Academy of Management Review* 12.1 (Jan. 1987), p. 9. ISSN: 03637425. DOI: 10.2307/257990.

[164] Jerald Greenberg. “Organizational Justice: Yesterday, Today, and Tomorrow”. In: *Journal of Management* 16.2 (June 1990), pp. 399–432. ISSN: 0149-2063. DOI: 10.1177/014920639001600208.

[165] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. Apr. 2018.

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

[167] Christopher Groves. “Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation”. In: *NanoEthics* 9.3 (Dec. 2015), pp. 321–333. ISSN: 1871-4757. DOI: 10.1007/s11569-015-0238-x.

[168] Siddharth Gulati, Sonia Sousa, and David Lamas. “Design, development and evaluation of a human-computer trust scale”. In: *Behaviour & Information Technology* 38.10 (Oct. 2019), pp. 1004–1015. ISSN: 0144-929X. DOI: 10.1080/0144929X.2019.1656779.

[169] Thilo Hagendorff. “The Ethics of Ai Ethics: An Evaluation of Guidelines”. In: *Minds and Machines* 30.1 (2020), pp. 99–120. DOI: 10.1007/s11023-020-09517-8.

[170] Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 3323–3331. ISBN: 9781510838819.

[171] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. “An empirical study on the perceived fairness of realistic, imperfect machine learning models”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2020, pp. 392–402. ISBN: 9781450369367. DOI: 10.1145/3351095.3372831.

[172] Michael R Harwell. “Choosing between parametric and nonparametric tests”. In: *Journal of Counseling & Development* 67.1 (1988), pp. 35–38.

[173] Tamarinde L. Haven and Dr. Leonie Van Grootel. “Preregistering qualitative research”. In: *Accountability in Research* 26 (3 Apr. 2019), pp. 229–244. ISSN: 0898-9621. DOI: 10.1080/08989621.2019.1580147.

[174] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. “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*. New York, NY, USA: ACM, Apr. 2023, pp. 1–18. ISBN: 9781450394215. DOI: 10.1145/3544548.3581025.

[175] Amy Heger, Elizabeth B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. “Understanding Machine Learning Practitioners’ Data Documentation Perceptions, Needs, Challenges, and Desiderata”. In: (June 2022).

[176] Katrina Heijne and Han van der Meer. *Road Map for Creative Problem Solving Techniques Organizing and facilitating group sessions*. Boom Uitgevers Amsterdam, Apr. 2019.

[177] Drew Hemment, Ruth Aylett, Vaishak Belle, Dave Murray-Rust, Ewa Luger, Jane Hillston, Michael Rovatsos, and Frank Broz. “Experiential AI”. In: *AI Matters* 5.1 (Apr. 2019), pp. 25–31. DOI: 10.1145/3320254.3320264. URL: https://doi.org/10.1145/3320254.3320264.

[178] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. “Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning”. In: (Jan. 2020).

[179] Clément Henin and Daniel Le Métayer. “Beyond explainability: justifiability and contestability of algorithmic decision systems”. In: *AI & SOCIETY* (July 2021). ISSN: 0951-5666. DOI: 10.1007/s00146-021-01251-8.

[180] Eelco Herder and Olaf van Maaren. “Privacy Dashboards: The Impact of the Type of Personal Data and User Control on Trust and Perceived Risk”. In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: ACM, July 2020, pp. 169–174. ISBN: 9781450379502. DOI: 10.1145/3386392.3399557.

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

[182] Mireille Hildebrandt. “Privacy as protection of the incomputable self: From agnostic to agonistic machine learning”. In: *Theoretical Inquiries in Law* 20.1 (2019), pp. 83–121.

[183] Mireille Hildebrandt. “Profiles and correlatable humans”. In: *Who Owns Knowledge?* Routledge, 2017, pp. 265–284.

[184] Simon David Hirsbrunner, Michael Tebbe, and Claudia Müller-Birn. "From critical technical practice to reflexive data science". In: *Convergence: The International Journal of Research into New Media Technologies* (Nov. 2022), p. 135485652211322. ISSN: 1354-8565. DOI: 10.1177/13548565221132243.

[185] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. "Designing Contestability". In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. New York, NY, USA: ACM, June 2017. ISBN: 9781450349222. DOI: 10.1145/3064663.3064703.

[186] Miriam Höddinghaus, Dominik Sondern, and Guido Hertel. "The automation of leadership functions: Would people trust decision algorithms?" In: *Computers in Human Behavior* 116 (Mar. 2021), p. 106635. ISSN: 07475632. DOI: 10.1016/j.chb.2020.106635.

[187] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards". In: (May 2018).

[188] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. "Improving Fairness in Machine Learning Systems". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2019, pp. 1–16. ISBN: 9781450359702. DOI: 10.1145/3290605.3300830.

6

[189] Leif-Erik Holtz, Katharina Nocun, and Marit Hansen. "Towards Displaying Privacy Information with Icons". In: 2011, pp. 338–348. DOI: 10.1007/978-3-642-20769-3{_}27.

[190] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. "Planning for Natural Language Failures with the AI Playbook". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–11. ISBN: 9781450380966. DOI: 10.1145/3411764.3445735.

[191] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. "Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. URL: https://doi.org/10.1145/3290605.3300637.

[192] Rhidian Hughes. "Considering the Vignette Technique and its Application to a Study of Drug Injecting and HIV Risk and Safer Behaviour". In: *Sociology of Health & Illness* 20.3 (May 1998), pp. 381–400. ISSN: 0141-9889. DOI: 10.1111/1467-9566.00107.

[193] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards Accountability for Machine Learning Datasets". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Mar. 2021, pp. 560–575. ISBN: 9781450383097. DOI: 10.1145/3442188.3445918.

[194] IBM. *IBM Everyday Ethics for AI*. 2019. URL: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

[195] IEEE. "IEEE Standard for Software Reviews and Audits". In: *IEEE Std 1028-2008* (2008), pp. 1–53. DOI: 10.1109/IEEESTD.2008.4601584.

[196] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. "An Agent-Based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 412–423. ISBN: 9781450383097. DOI: 10.1145/3442188.3445904. URL: https://doi.org/10.1145/3442188.3445904.

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

[198] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection". In: *Translational Psychiatry* 11 (1 Feb. 2021), p. 108. ISSN: 2158-3188. DOI: 10.1038/s41398-021-01224-x.

[199] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning". In: *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2018, pp. 19–35.

[200] Marijn Janssen, Martijn Hartog, Ricardo Matheus, Aaron Yi Ding, and George Kuk. "Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government". In: *Social Science Computer Review* 40.2 (Apr. 2022), pp. 478–493. ISSN: 0894-4393. DOI: 10.1177/0894439320980118.

[201] Technology Japanese Cabinet Office Council for Science and Innovation. *Social Principles of Human-Centric Artificial Intelligence*. 2019. URL: https://www8.cao.go.jp/cstp/english/humancentricai.pdf.

[202] Harold Jeffreys. "Theory of Probability. (1939)". In: (1939).

[203] Denise Jepsen and John Rodwell. "A New Dimension of Organizational Justice: Procedural Voice". In: *Psychological Reports* 105.2 (Oct. 2009), pp. 411–426. ISSN: 0033-2941. DOI: 10.2466/PR0.105.2.411-426.

[204] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. "MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples". In: (Sept. 2019).

[205] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. "EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence". In: (Feb. 2021). URL: https://arxiv.org/abs/2102.02437.

[206] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. “Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems”. In: (July 2019).

[207] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. “Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion”. In: *Proceedings of the 28th European Conference on Information Systems (ECIS)*. 2020.

[208] Jonas Kaiser and Adrian Rauchfleisch. “Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube’s Channel Recommendations in the United States and Germany”. In: *Social Media + Society* 6.4 (Oct. 2020), p. 2056305120969914. ISSN: 2056-3051. DOI: 10.1177/2056305120969914. URL: https://doi.org/10.1177/2056305120969914.

[209] Pratyusha Kalluri. “Don’t ask if artificial intelligence is good or fair, ask how it shifts power.” In: *Nature* 583.7815 (2020). ISSN: 1476-4687. DOI: 10.1038/d41586-020-02003-2.

[210] Margot E Kaminski and Jennifer M Urban. “THE RIGHT TO CONTEST AI”. In: *Columbia Law Review* 121.7 (2021), pp. 1957–2048. ISSN: 00101958, 19452268. URL: https://www-jstor-org.tudelft.idm.oclc.org/stable/27083420.

[211] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. “”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*. New York, NY, USA: ACM, Apr. 2022, pp. 1–18. ISBN: 9781450391573. DOI: 10.1145/3491102.3517533.

[212] Andreas Kaplan and Michael Haenlein. “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence”. In: *Business horizons* 62.1 (2019), pp. 15–25.

[213] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. “A survey of algorithmic recourse:contrastive explanations and consequential recommendations”. In: *ACM Computing Surveys* (Apr. 2022). ISSN: 0360-0300. DOI: 10.1145/3527848.

[214] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. “Algorithmic Recourse”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Mar. 2021, pp. 353–362. ISBN: 9781450383097. DOI: 10.1145/3442188.3445899.

[215] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. “On the Effectiveness of Regularization Against Membership Inference Attacks”. In: (June 2020).

[216] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Nov. 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

[217] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* USA: Oxford University Press, Inc., 2019. ISBN: 0190948205.

[218] Os Keyes, Jevan Hutson, and Meredith Durbin. "A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* CHI EA '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–11. ISBN: 9781450359719. DOI: 10.1145/3290607.3310433. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3290607.3310433.

[219] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. ""Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM, Apr. 2023, pp. 1–17. ISBN: 9781450394215. DOI: 10.1145/3544548.3581001.

[220] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application". In: *2023 ACM Conference on Fairness, Accountability, and Transparency.* New York, NY, USA: ACM, June 2023, pp. 77–88. ISBN: 9798400701924. DOI: 10.1145/3593013.3593978.

[221] Nigel King. "Doing template analysis". In: *Qualitative organizational research: Core methods and current challenges* 426 (2012), pp. 426–450.

[222] Styliani Kleanthous, Maria Kasinidou, Pınar Barlas, and Jahna Otterbacher. "Perception of fairness in algorithmic decisions: Future developers' perspective". In: *Patterns* 3.1 (Jan. 2022), p. 100380. ISSN: 26663899. DOI: 10.1016/j.patter.2021.100380.

[223] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: (Sept. 2016).

[224] Daniel Kluttz, Nitin Kohli, and Deirdre K. Mulligan. "Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems". In: *SSRN Electronic Journal* (2018). ISSN: 1556-5068. DOI: 10.2139/ssrn.3311894.

[225] Daniel Kluttz and Deirdre K. Mulligan. "Automated decision support technologies and the legal profession". In: *SSRN Electronic Journal* (2019). ISSN: 1556-5068. DOI: 10.2139/ssrn.3443063.

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

[227] kobi leins kobi, Jey Han Lau, and Timothy Baldwin. "Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.261.

[228] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. “Will You Accept an Imperfect AI?” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2019, pp. 1–14. ISBN: 9781450359702. DOI: 10.1145/3290605.3300641.

[229] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. “Measurement of Trust in Automation: A Narrative Review and Reference Guide”. In: *Frontiers in Psychology* 12 (Oct. 2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.604977.

[230] Daan Kolkman. “The usefulness of algorithmic models in policy making”. In: *Government Information Quarterly* 37.3 (July 2020), p. 101488. ISSN: 0740624X. DOI: 10.1016/j.giq.2020.101488.

[231] Sherrie Yi Xiao Komiak. “The impact of internalization and familiarity on trust and adoption of recommendation agents”. eng. PhD thesis. 2003. URL: https://open.library.ubc.ca/collections/831/items/1.0091325.

[232] TD Krafft and K Zweig. “Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse”. In: *Ein Regulierungsvorschlag* (2019).

[233] Max F. Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. “When Do People Want AI to Make Decisions?” In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, Dec. 2018, pp. 204–209. ISBN: 9781450360128. DOI: 10.1145/3278721.3278752.

[234] Lenneke Kuijer and Elisa Giaccardi. “Co-Performance: Conceptualizing the Role of Artificial Agency in the Design of Everyday Life”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. URL: https://doi.org/10.1145/3173574.3173699.

[235] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. “Principles of Explanatory Debugging to Personalize Interactive Machine Learning”. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, Mar. 2015, pp. 126–137. ISBN: 9781450333061. DOI: 10.1145/2678025.2701399.

[236] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. “POTs: Protective Optimization Technologies”. In: (June 2018). DOI: 10.1145/3351095.3372853.

[237] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. “Counterfactual Fairness”. In: *Advances in Neural Information Processing Systems*. Ed. by I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[238] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. “End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (Nov. 2022), pp. 1–34. ISSN: 2573-0142. DOI: 10.1145/3555625.

[239] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. ""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*. New York, NY, USA: ACM, Apr. 2022, pp. 1–28. ISBN: 9781450391573. DOI: 10.1145/3491102.3517527.

[240] Claire Larsonneur. *Intelligence artificielle ET/OU diversité linguistique : les paradoxes du traitement automatique des langues*. 2021. URL: http://www.hybrid.univ-paris8.fr/lodel/index.php?id=1542.

[241] Douglass B. Lee. "Requiem for Large-Scale Models". In: *Journal of the American Institute of Planners* 39.3 (May 1973), pp. 163–178. ISSN: 0002-8991. DOI: 10.1080/01944367308977851.

[242] J. D. Lee and K. A. See. "Trust in Automation: Designing for Appropriate Reliance". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (Jan. 2004). ISSN: 0018-7208. DOI: 10.1518/hfes.46.1.50{_}30392.

[243] Michael D. Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling*. Cambridge University Press, Apr. 2014. ISBN: 9781107603578. DOI: 10.1017/CBO9781139087759.

[244] Michelle Seng Ah Lee and Jat Singh. "The Landscape and Gaps in Open Source Fairness Toolkits". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–13. ISBN: 9781450380966. DOI: 10.1145/3411764.3445261.

[245] Michelle Seng Ah Lee and Jatinder Singh. "Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, July 2021, pp. 704–714. ISBN: 9781450384735. DOI: 10.1145/3461702.3462572.

[246] Min Hun Lee and Chong Jun Chew. "Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW2 (Oct. 2023). DOI: 10.1145/3610218. URL: https://doi.org/10.1145/3610218.

[247] Min Kyung Lee. "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management". In: *Big Data & Society* 5.1 (Jan. 2018). ISSN: 2053-9517. DOI: 10.1177/2053951718756684.

[248] Min Kyung Lee and Su Baykal. "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1035–1048. ISBN: 9781450343350. DOI: 10.1145/2998181.2998230. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2998181.2998230.

[249] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. "Procedural Justice in Algorithmic Fairness". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–26. ISSN: 2573-0142. DOI: 10.1145/3359284.

[250] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D Procaccia. "WeBuildAI: Participatory Framework for Algorithmic Governance". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359283. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3359283.

[251] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. "WeBuildAI". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–35. ISSN: 2573-0142. DOI: 10.1145/3359283.

[252] Min Kyung Lee and Katherine Rich. "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*. New York, NY, USA: ACM, May 2021, pp. 1–14. ISBN: 9781450380966. DOI: 10.1145/3411764.3445570.

[253] Gerald S. Leventhal. "What Should Be Done with Equity Theory?" In: *Social Exchange*. Boston, MA: Springer US, 1980, pp. 27–55. DOI: 10.1007/978-1-4613-3087-5{_}2.

[254] Roy J Lewicki and Barbara Benedict Bunker. *Trust in relationships: A model of development and decline*. Jossey-Bass/Wiley, 1995. ISBN: 0787900699.

[255] Q. Vera Liao, Daniel Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences". In: (Jan. 2020). DOI: 10.1145/3313831.3376590. URL: https://arxiv.org/abs/2001.02478.

[256] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. "Question-Driven Design Process for Explainable AI User Experiences". In: (Apr. 2021).

[257] Q. Vera Liao and S. Shyam Sundar. "Designing for Responsible Trust in AI Systems: A Communication Perspective". In: (Apr. 2022). DOI: 10.1145/3531146.3533182.

[258] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. "Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10.1 (Oct. 2022), pp. 147–159. ISSN: 2769-1349. DOI: 10.1609/hcomp.v10i1.21995.

[259] E. Allan Lind and Tom R. Tyler. *The Social Psychology of Procedural Justice*. Boston, MA: Springer US, 1988. ISBN: 978-1-4899-2117-8. DOI: 10.1007/978-1-4899-2115-4.

[260] Charles Lindblom. "The science of "muddling through"". In: *Classic readings in urban planning*. Routledge, 2018, pp. 31–40.

[261] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. "Axies: Identifying and Evaluating Context-Specific Values". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2021, pp. 799–808. ISBN: 9781450383073.

[262] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. "Algorithm appreciation: People prefer algorithmic to human judgment". In: *Organizational Behavior and Human Decision Processes* 151 (Mar. 2019), pp. 90–103. ISSN: 07495978. DOI: 10.1016/j.obhdp.2018.12.005.

[263] Duri Long and Brian Magerko. "What is AI Literacy? Competencies and Design Considerations". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–16. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376727.

[264] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. "Resistance to Medical Artificial Intelligence". In: *Journal of Consumer Research* 46.4 (Dec. 2019), pp. 629–650. ISSN: 0093-5301. DOI: 10.1093/jcr/ucz013.

[265] Christopher T Lowenkamp. "The development of an actuarial risk assessment instrument for US Pretrial Services". In: *Fed. Probation* 73 (2009), p. 33.

[266] Sasha Luccioni, Boris Gamazaychikov, Sara Hooker, Régis Pierrard, Emma Strubell, Yacine Jernite, and Carole-Jean Wu. "Light bulbs have energy ratings—so why can't AI chatbots?" In: *Nature* 632.8026 (2024), pp. 736–738.

[267] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[268] Henrietta Lyons, Tim Miller, and Eduardo Velloso. "Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review". In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2023, pp. 764–774. ISBN: 9798400701924. DOI: 10.1145/3593013.3594041.

[269] Henrietta Lyons, Eduardo Velloso, and Tim Miller. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions". In: (Feb. 2021). DOI: 10.1145/3449180.

[270] Henrietta Lyons, Eduardo Velloso, and Tim Miller. "Designing for Contestation: Insights from Administrative Law". In: *2019 Workshop on Contestability in Algorithmic Systems at CSCW Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '19)*. Feb. 2019.

[271] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. "What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions". In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–15. ISBN: 9781450391573. DOI: 10.1145/3491102.3517606.

[272] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. “A Unified Gradient Regularization Family for Adversarial Examples”. In: *2015 IEEE International Conference on Data Mining*. IEEE, Nov. 2015, pp. 301–309. ISBN: 978-1-4673-9504-5. DOI: 10.1109/ICDM.2015.84.

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

[274] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. “Assessing the Fairness of AI Systems: AI Practitioners’ Processes, Challenges, and Needs for Support”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1 (Mar. 2022), pp. 1–26. ISSN: 2573-0142. DOI: 10.1145/3512899.

[275] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376445.

[276] Anna Madill, Abbie Jordan, and Caroline Shirley. “Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies”. In: *British Journal of Psychology* 91.1 (Feb. 2000), pp. 1–20. ISSN: 00071269. DOI: 10.1348/000712600161646.

[277] N. Mahendran. “Analysis of memory consumption by neural networks based on hyperparameters”. In: (Oct. 2021).

[278] Donald Martin, Jr Google Vinodkumar Prabhakaran Google Jill Kuhlberg, and Andrew S Smart Google William Isaac DeepMind. “Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context”. In: (2020).

[279] Roger C. Mayer and James H. Davis. “The effect of the performance appraisal system on trust for management: A field quasi-experiment.” In: *Journal of Applied Psychology* 84.1 (Feb. 1999), pp. 123–136. ISSN: 1939-1854. DOI: 10.1037/0021-9010.84.1.123.

[280] Roger C. Mayer, James H. Davis, and F. David Schoorman. “An Integrative Model of Organizational Trust”. In: *The Academy of Management Review* 20.3 (July 1995), p. 709. ISSN: 03637425. DOI: 10.2307/258792.

[281] D. J. McAllister. “AFFECT- AND COGNITION-BASED TRUST AS FOUNDATIONS FOR INTERPERSONAL COOPERATION IN ORGANIZATIONS.” In: *Academy of Management Journal* 38.1 (Feb. 1995), pp. 24–59. ISSN: 0001-4273. DOI: 10.2307/256727.

[282] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. “Developing and Validating Trust Measures for e-Commerce: An Integrative Typology”. In: *Information Systems Research* 13.3 (Sept. 2002), pp. 334–359. ISSN: 1047-7047. DOI: 10.1287/isre.13.3.334.81.

[283] M. Mehldau. *Iconset for data-privacy declarations v 0.1.* 2007. URL: https://netzpolitik.org/wp-upload/data-privacy-icons-v01.pdf.

[284] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Comput. Surv.* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3457607.

[285] Albert Meijer and Martijn Wessels. “Predictive Policing: Review of Benefits and Drawbacks.” In: 42 (Dec. 2019), pp. 1031–1039. URL: https://doi.org/10.1080/01900692.2019.1575664.

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

[287] Isak Mendoza and Lee A. Bygrave. “The Right Not to be Subject to Automated Decisions Based on Profiling”. In: *EU Internet Law: Regulation and Enforcement.* Ed. by Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Markou, and Thalia Prastitou. Cham: Springer International Publishing, 2017, pp. 77–98. ISBN: 978-3-319-64955-9. DOI: 10.1007/978-3-319-64955-9_4. URL: https://doi.org/10.1007/978-3-319-64955-9_4.

[288] Ryan Merrill and Nicole Sintov. “An Affinity-to-Commons Model of Public Support For Environmental Energy Policy”. In: *Energy Policy* 99 (Dec. 2016), pp. 88–99. ISSN: 03014215. DOI: 10.1016/j.enpol.2016.09.048.

[289] Microsoft. *AI Principles.* 2018. URL: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6.

[290] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38. ISSN: 00043702. DOI: 10.1016/j.artint.2018.07.007.

[291] Swati Mishra and Jeffrey M Rzeszotarski. “Designing Interactive Transfer Learning Tools for ML Non-Experts”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM, May 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445096.

[292] Mission assigned by the French Prime Minister. *For a Meaningful Artificial Intelligence: Toward a French and European Strategy.* 2019. URL: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

[293] Gianluca Misuraca, Colin van Noordt, and Anys Boukli. “The use of AI in public services: Results from a preliminary mapping across the EU”. In: 2020, pp. 90–99. URL: https://doi.org/10.1145/3428502.3428513.

[294] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting”. In: (Oct. 2018). DOI: 10.1145/3287560.3287596.

[295] Tanushree Mitra. “Provocation: Contestability in Large-Scale Interactive {NLP} Systems”. In: *Proceedings of the First Workshop on Bridging Human{–}Computer Interaction and Natural Language Processing.* Association for Computational Linguistics, Apr. 2021, pp. 96–100.

[296] Brent Mittelstadt. “Principles alone cannot guarantee ethical AI”. In: *Nature Machine Intelligence* 1.11 (2019), pp. 501–507. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0114-4. URL: https://doi.org/10.1038/s42256-019-0114-4.

[297] Jakub Mlynar, Farzaneh Bahrami, André Ourednik, Nico Mutzner, Himanshu Verma, and Hamed Alavi. “AI beyond Deus ex Machina – Reimagining Intelligence in Future Cities with Urban Experts”. In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–13. ISBN: 9781450391573. DOI: 10.1145/3491102.3517502.

[298] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices”. In: *Science and Engineering Ethics* 26 (2020), pp. 2141–2168. DOI: 10.1007/s11948-019-00165-5. URL: https://doi.org/10.1007/s11948-019-00165-5.

[299] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. In: (May 2019). DOI: 10.1145/3351095.3372850.

[300] Michael Muller and Angelika Strohmayer. “Forgetting Practices in the Data Sciences”. In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–19. ISBN: 9781450391573. DOI: 10.1145/3491102.3517644.

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

[302] Pradeep K Murukannaiah and Munindar P Singh. “Xipho: Extending Tropos to Engineer Context-Aware Personal Agents”. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. AAMAS ’14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 309–316. ISBN: 9781450327381.

[303] Rosanna Nagtegaal. “The impact of using algorithms for managerial decisions on public employees’ procedural justice”. In: *Government Information Quarterly* 38.1 (Jan. 2021), p. 101536. ISSN: 0740624X. DOI: 10.1016/j.giq.2020.101536.

[304] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. “Towards Involving End-users in Interactive Human-in-the-loop AI Fairness”. In: (Apr. 2022).

[305] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Machine Learning with Membership Privacy using Adversarial Regularization”. In: (July 2018).

[306] DCODE Network. *DCODE Network*. 2024. URL: https://dcode-network.eu/.

[307] David T. Newman, Nathanael J. Fast, and Derek J. Harmon. “When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions”. In: *Organizational Behavior and Human Decision Processes* 160 (Sept. 2020), pp. 149–167. ISSN: 07495978. DOI: 10.1016/j.obhdp.2020.03.008.

[308] Shirley Nieuwland and Rianne van Melik. “Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals”. In: *Current Issues in Tourism* 23.7 (Apr. 2020), pp. 811–825. ISSN: 1368-3500. DOI: 10.1080/13683500.2018.1504899.

[309] Colin van Noordt and Gianluca Misuraca. “Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union”. In: *Government Information Quarterly* 39 (3 2022). URL: https://doi.org/10.1016/j.giq.2022.101714.

[310] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. “InterpretML: A Unified Framework for Machine Learning Interpretability”. In: (Sept. 2019).

[311] Arif Nurwidyantoro, Mojtaba Shahin, Michel Chaudron, Waqar Hussain, Harsha Perera, Rifat Ara Shams, and Jon Whittle. “Towards a Human Values Dashboard for Software Development: An Exploratory Study”. In: (July 2021).

[312] Ronald C. Nyhan and Herbert A. Marlowe. “Development and Psychometric Properties of the Organizational Trust Inventory”. In: *Evaluation Review* 21.5 (Oct. 1997), pp. 614–635. ISSN: 0193-841X. DOI: 10.1177/0193841X9702100505.

[313] Kieron O’Hara. “Explainable AI and the philosophy and practice of explanation”. In: *Computer Law & Security Review* 39 (Nov. 2020), p. 105474. ISSN: 0267-3649. DOI: 10.1016/J.CLSR.2020.105474.

[314] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2016.

[315] OECD. *Recommendation of the Council on Artificial Intelligence.* 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0406.

[316] Gideon Ogunniye, Benedicte Legastelois, Michael Rovatsos, Liz Dowthwaite, Virginia Portillo, Elvira Perez Vallejos, Jun Zhao, and Marina Jirotka. “Understanding User Perceptions of Trustworthiness in E-Recruitment Systems”. In: *IEEE Internet Computing* 25.6 (Nov. 2021), pp. 23–32. ISSN: 1089-7801. DOI: 10.1109/MIC.2021.3115670.

[317] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. “Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1 (Mar. 2022), pp. 1–31. ISSN: 2573-0142. DOI: 10.1145/3512929.

[318] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. “Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems”. In: *CHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM, Apr. 2022, pp. 1–9. ISBN: 9781450391573. DOI: 10.1145/3491102.3502104.

[319] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. “Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks”. In: *2016 IEEE Symposium on Security and Privacy (SP).* IEEE, May 2016, pp. 582–597. ISBN: 978-1-5090-0824-7. DOI: 10.1109/SP.2016.41.

[320] Lorenza Parisi and Francesca Comunello. “Dating in the time of “relational filter bubbles”: exploring imaginaries, perceptions and tactics of Italian dating app users”. In: *The Communication Review* 23.1 (2020), pp. 66–89. DOI: 10.1080/10714421.2019.1704111. URL: https://doi.org/10.1080/10714421.2019.1704111.

[321] Samir Passi and Solon Barocas. “Problem Formulation and Fairness”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2019, pp. 39–48. ISBN: 9781450361255. DOI: 10.1145/3287560.3287567.

[322] Samir Passi and Steven J. Jackson. “Trust in Data Science”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018), pp. 1–28. ISSN: 2573-0142. DOI: 10.1145/3274405.

[323] Samir Passi and Phoebe Sengers. “Making data science systems work”. In: *Big Data & Society* 7.2 (July 2020), p. 205395172093960. ISSN: 2053-9517. DOI: 10.1177/2053951720939605.

[324] Reema Patel. “Reboot AI with human values”. In: *Nature* 598.7879 (Oct. 2021). ISSN: 0028-0836. DOI: 10.1038/d41586-021-02693-2.

[325] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. “Data and its (dis)contents: A survey of dataset development and use in machine learning research”. In: (Dec. 2020). URL: http://arxiv.org/abs/2012.05345.

[326] Virgil Pavlu, Shahzad Rajput, Peter B. Golbus, and Javed A. Aslam. “IR system evaluation using nugget-based test collections”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, Feb. 2012, pp. 393–402. ISBN: 9781450307475. DOI: 10.1145/2124295.2124343.

[327] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. “What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 125–134. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5281.

[328] Kathleen H. Pine and Max Liboiron. “The Politics of Measurement and Action”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2015, pp. 3147–3156. ISBN: 9781450331456. DOI: 10.1145/2702123.2702298.

[329] Ibo van de Poel. “Translating Values into Design Requirements”. In: 2013, pp. 253–266. DOI: 10.1007/978-94-007-7762-0{_}20.

[330] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn Jonker. “Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate”. In: *Ethics and Information Technology* 14.4 (Dec. 2012), pp. 285–303. ISSN: 1388-1957. DOI: 10.1007/s10676-011-9282-6.

[331] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–52. ISBN: 9781450380966. DOI: 10.1145/3411764.3445315.

[332] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 469–481. ISBN: 9781450369367. DOI: 10.1145/3351095.3372828. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3351095.3372828.

[333] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. “The Fallacy of AI Functionality”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 959–972. ISBN: 9781450393522. DOI: 10.1145/3531146.3533158.

[334] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. “Closing the AI accountability gap”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2020, pp. 33–44. ISBN: 9781450369367. DOI: 10.1145/3351095.3372873.

[335] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. “Where Responsible AI meets Reality”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021), pp. 1–23. ISSN: 2573-0142. DOI: 10.1145/3449081.

[336] Peyman Rasouli and Ingrid Chieh Yu. “CARE: coherent actionable recourse based on sound counterfactual explanations”. In: *International Journal of Data Science and Analytics* (Sept. 2022). ISSN: 2364-415X. DOI: 10.1007/s41060-022-00365-6.

[337] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A F Almeida, and Wagner Meira. “Auditing Radicalization Pathways on YouTube”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 131–141. ISBN: 9781450369367. DOI: 10.1145/3351095.3372879. URL: https://doi.org/10.1145/3351095.3372879.

[338] De rijksoverheid. *Strategisch Actieplan voor Artificiële Intelligentie*. 2019. URL: https://open.overheid.nl/documenten/ronl-e14cdcee-690c-4995-9870-fa4141319d6f/pdf.

[339] Jeanne S Ringel, Dana Schultz, Joshua Mendelsohn, Stephanie Brooks Holliday, Katharine Sieck, Ifeanyi Edochie, and Lauren Davis. “Improving child welfare outcomes: balancing investments in prevention and treatment”. In: *Rand health quarterly* 7.4 (2018).

[340] Horst W. J. Rittel and Melvin M. Webber. “Dilemmas in a general theory of planning”. In: *Policy Sciences* 4.2 (June 1973), pp. 155–169. ISSN: 0032-2687. DOI: 10.1007/BF01405730.

[341] Karlene H Roberts and Charles A O'Reilly. "Measuring organizational communication." In: *Journal of applied psychology* 59.3 (1974), p. 321. ISSN: 1939-1854.

[342] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. "Tackling climate change with machine learning". In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–96.

[343] Arianna Rossi and Monica Palmirani. "A Visualization Approach for Adaptive Consent in the European Data Protection Framework". In: *2017 Conference for E-Democracy and Open Government (CeDEM)*. IEEE, May 2017, pp. 159–170. ISBN: 978-1-5090-6718-3. DOI: 10.1109/CeDEM.2017.23.

[344] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. "Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust". In: *The Academy of Management Review* 23.3 (1998), pp. 393–404. ISSN: 03637425. URL: http://www.jstor.org/stable/259285.

[345] Alan Rubel, Clinton Castro, and Adam Pham. *Algorithms and Autonomy*. Cambridge University Press, Apr. 2021. ISBN: 9781108895057. DOI: 10.1017/9781108895057.

[346] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[347] Stuart Russell, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence". In: *AI Magazine* 36.4 (Dec. 2015). ISSN: 2371-9621. DOI: 10.1609/aimag.v36i4.2577.

[348] Shadan Sadeghian, Alarith Uhde, and Marc Hassenzahl. "The Soul of Work: Evaluation of Job Meaningfulness and Accountability in Human-AI Collaboration". In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). DOI: 10.1145/3637407. URL: https://doi.org/10.1145/3637407.

[349] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. "Aequitas: A Bias and Fairness Audit Toolkit". In: 2018.

[350] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. ""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*. New York, NY, USA: ACM, May 2021, pp. 1–15. ISBN: 9781450380966. DOI: 10.1145/3411764.3445518.

[351] Helen Sampson and Idar Alfred Johannessen. "Turning on the tap: the benefits of using 'real-life' vignettes in qualitative research interviews". In: *Qualitative Research* 20.1 (Feb. 2020), pp. 56–72. ISSN: 1468-7941. DOI: 10.1177/1468794118816618.

[352] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." In: *Data and discrimination: converting critical concerns into productive inquiry 22*. 2014.

[353] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. "The Risk of Racial Bias in Hate Speech Detection". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (2019), pp. 1668–1678. DOI: 10.18653/V1/P19-1163. URL: https://aclanthology.org/P19-1163.

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

[355] Claudio Sarra. "Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of Contestability by Design". In: *Global Jurist* 20.3 (Oct. 2020). ISSN: 1934-2640. DOI: 10.1515/gj-2020-0003.

[356] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–41. ISSN: 2573-0142. DOI: 10.1145/3476089.

[357] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. "How Do Fairness Definitions Fare?" In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, Jan. 2019, pp. 99–106. ISBN: 9781450363242. DOI: 10.1145/3306618.3314248.

[358] Enrique Schaerer, Richard Kelley, and Monica Nicolescu. "Robots as animals: A framework for liability and responsibility in human-robot interactions". In: *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Sept. 2009. ISBN: 978-1-4244-5081-7. DOI: 10.1109/ROMAN.2009.5326244.

[359] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. "FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions". In: (Nov. 2019).

[360] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development". In: *Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 317* (2021). URL: https://doi.org/10.1145/3476058.

[361] Philipp Schmidt and Felix Biessmann. "Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias". In: 2020, pp. 431–449. DOI: 10.1007/978-3-030-57321-8{_}24.

[362] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. “On the Impact of Explanations on Understanding of Algorithmic Decision-Making”. In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2023, pp. 959–970. ISBN: 9798400701924. DOI: 10.1145/3593013.3594054.

[363] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. “Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642621. URL: https://doi.org/10.1145/3613904.3642621.

[364] Jakob Schoeffer and Niklas Kuehl. “Appropriate Fairness Perceptions? On the Effectiveness of Explanations in Enabling People to Assess the Fairness of Automated Decision Systems”. In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. New York, NY, USA: ACM, Oct. 2021, pp. 153–157. ISBN: 9781450384797. DOI: 10.1145/3462204.3481742.

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

[366] Shalom H. Schwartz. “An Overview of the Schwartz Theory of Basic Values”. In: *Online Readings in Psychology and Culture* 2.1 (Dec. 2012). ISSN: 2307-0919. DOI: 10.9707/2307-0919.1116.

[367] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. “Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 2138–2148. ISBN: 9781450393522. DOI: 10.1145/3531146.3534631.

[368] Andrew D Selbst and Julia Powles. “Meaningful information and the right to explanation”. In: *International Data Privacy Law* 7.4 (Nov. 2017), pp. 233–242. ISSN: 2044-3994. DOI: 10.1093/idpl/ipx022.

[369] Mojtaba Shahin, Waqar Hussain, Arif Nurwidyantoro, Harsha Perera, Rifat Shams, John Grundy, and Jon Whittle. “Operationalizing Human Values in Software Engineering: A Survey”. In: (Aug. 2021).

[370] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. “Why Am I Not Seeing It? Understanding Users’ Needs for Counterfactual Explanations in Everyday Recommendations”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 1330–1340. ISBN: 9781450393522. DOI: 10.1145/3531146.3533189.

[371] Debra L. Shapiro, E.Holly Buttner, and Bruce Barry. “Explanations: What Factors Enhance Their Perceived Adequacy?” In: *Organizational Behavior and Human Decision Processes* 58.3 (June 1994), pp. 346–368. ISSN: 07495978. DOI: 10.1006/obhd.1994.1041.

[372] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. “Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors”. In: (May 2021). DOI: 10.1145/3479577.

[373] Irina Shklovski and Carolina Némethy. “Nodes of certainty and spaces for doubt in AI ethics for engineers”. In: *Information, Communication & Society* (Jan. 2022), pp. 1–17. ISSN: 1369-118X. DOI: 10.1080/1369118X.2021.2014547.

[374] Ben Shneiderman. “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems”. In: *ACM Trans. Interact. Intell. Syst.* 10.4 (Oct. 2020). ISSN: 2160-6455. DOI: 10.1145/3419764. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3419764.

[375] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. “Membership Inference Attacks against Machine Learning Models”. In: (Oct. 2016).

[376] Herbert A Simon. “Bounded rationality”. In: *Utility and probability* (1990), pp. 15–18. ISSN: 0333495411.

[377] Jesper Simonsen and Toni Robertson. *Routledge international handbook of participatory design*. Vol. 711. Routledge New York, 2013.

[378] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. “Directive Explanations for Actionable Explainability in Machine Learning Applications”. In: *ACM Transactions on Interactive Intelligent Systems* (Jan. 2023). ISSN: 2160-6455. DOI: 10.1145/3579363.

[379] Kacper Sokol and Peter Flach. “Explainability fact sheets”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2020. ISBN: 9781450369367. DOI: 10.1145/3351095.3372870.

[380] Elizabeth Solberg, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. “A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids”. In: *Group & Organization Management* 47.2 (Apr. 2022), pp. 187–222. ISSN: 1059-6011. DOI: 10.1177/10596011221081238.

[381] Megha Srivastava, Hoda Heidari, and Andreas Krause. “Mathematical Notions vs. Human Perception of Fairness”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: ACM, July 2019, pp. 2459–2468. ISBN: 9781450362016. DOI: 10.1145/3292500.3330664.

[382] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. “Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature”. In: *Big Data & Society* 9.2 (July 2022), p. 205395172211151. ISSN: 2053-9517. DOI: 10.1177/20539517221115189.

[383] H. Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. “Social transparency in networked information exchange”. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, Feb. 2012, pp. 451–460. ISBN: 9781450310864. DOI: 10.1145/2145204.2145275.

[384] Emily Sullivan and Philippe Verreault-Julien. “From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, July 2022, pp. 712–722. ISBN: 9781450392471. DOI: 10.1145/3514094.3534185.

[385] Tara Qian Sun and Rony Medaglia. “Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare”. In: *Government Information Quarterly* 36.2 (Apr. 2019), pp. 368–383. ISSN: 0740-624X. DOI: 10.1016/J.GIQ.2018.09.008.

[386] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. “Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, May 2021, pp. 1–16. ISBN: 9781450380966. DOI: 10.1145/3411764.3445088.

[387] Harini Suresh and John Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle”. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: ACM, Oct. 2021, pp. 1–9. ISBN: 9781450385534. DOI: 10.1145/3465416.3483305.

[388] Jacob Kraemer Tebes. “Community science, philosophy of science, and the practice of research”. In: *American journal of community psychology* 35.3-4 (2005), pp. 213–230. ISSN: 1573-2770.

[389] Telia Company. *Guiding Principles on Trusted AI Ethics*. 2019. URL: https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf.

[390] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. First Edition. IEEE, 2019.

[391] The Royal Society. *Explainable AI: the basics*. 2019. URL: https://doi.org/10.1177/1461444816676645.

[392] Sarah Thew and Alistair Sutcliffe. “Value-based requirements engineering: method and experience”. In: *Requirements Engineering* 23.4 (Nov. 2018). ISSN: 0947-3602. DOI: 10.1007/s00766-017-0273-y.

[393] J. W. Thibaut and L. Walker. “Procedural Justice: A Psychological Analysis”. In: *L. Erlbaum Associates, Hillsdale.* (1975).

[394] Nava Tintarev and Judith Masthoff. “Effective explanations of recommendations”. In: *Proceedings of the 2007 ACM conference on Recommender systems*. New York, NY, USA: ACM, Oct. 2007, pp. 153–156. ISBN: 9781595937308. DOI: 10.1145/1297231.1297259.

[395] Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. “A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities”. In: *ACM Comput. Surv.* (May 2024). Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3665926. URL: https://doi.org/10.1145/3665926.

6

[396] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. "Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ICAIL '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 83–92. ISBN: 9781450367547. DOI: 10.1145/3322640.3326705. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3322640.3326705.

[397] Dimitrios Tsarapatsanis and Nikolaos Aletras. "On the Ethical Limits of Natural Language Processing on Legal Text". In: (May 2021).

[398] Tom R Tyler. "The psychology of procedural justice: A test of the group-value model." In: *Journal of personality and social psychology* 57.5 (1989), p. 830.

[399] Tom R. Tyler. "Procedural justice, legitimacy, and the effective rule of law." In: *Crime and justice* 30 (2003), pp. 283–357.

[400] Tom R. Tyler. "What is Procedural Justice?: Criteria used by Citizens to Assess the Fairness of Legal Procedures". In: *Law & Society Review* 22.1 (1988), p. 103. ISSN: 00239216. DOI: 10.2307/3053563.

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

[402] National Science United States Executive Office of the President and Technology Council Committee on Technology. *Preparing for the Future of Artificial Intelligence*. 2016. URL: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

[403] Funda Ustek-Spilda, Alison Powell, and Selena Nemorin. "Engaging with ethics in Internet of Things: Imaginaries in the social milieu of technology developers". In: *Big Data & Society* 6.2 (July 2019), p. 205395171987946. ISSN: 2053-9517. DOI: 10.1177/2053951719879468.

[404] Berk Ustun, Alexander Spangher, and Yang Liu. "Actionable Recourse in Linear Classification". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2019, pp. 10–19. ISBN: 9781450361255. DOI: 10.1145/3287560.3287566.

[405] Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. "Contestability in Algorithmic Systems". In: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. New York, NY, USA: ACM, Nov. 2019, pp. 523–527. ISBN: 9781450366922. DOI: 10.1145/3311957.3359435.

[406] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. ""At the End of the Day Facebook Does What ItWants"". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (Oct. 2020), pp. 1–22. ISSN: 2573-0142. DOI: 10.1145/3415238.

[407] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. “Contestability For Content Moderation”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–28. ISSN: 2573-0142. DOI: 10.1145/347605 9.

[408] Orlin Vakarelov and Kenneth Rogerson. “The Transparency Game: Government Information, Access, and Actionability”. In: *Philosophy & Technology* 33.1 (Mar. 2020), pp. 71–92. ISSN: 2210-5433. DOI: 10.1007/s13347-019-0340-z.

[409] Annukka Valkeapää and Tuija Seppälä. “Speed of Decision-Making as a Procedural Justice Principle”. In: *Social Justice Research* 27.3 (Sept. 2014), pp. 305–321. ISSN: 0885-7466. DOI: 10.1007/s11211-014-0214-6.

[410] Arnaud Van Looveren and Janis Klaise. “Interpretable Counterfactual Explanations Guided by Prototypes”. In: 2021, pp. 650–665. DOI: 10.1007/978-3-030-865 20-7{_}40.

[411] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making”. In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (Apr. 2023), pp. 1–38. ISSN: 2573-0142. DOI: 10.1145/3579605.

[412] Suresh Venkatasubramanian and Mark Alfano. “The philosophical basis of algorithmic recourse”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jan. 2020, pp. 284–293. ISBN: 9781450369367. DOI: 10.1145/3351095.3372876.

[413] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. “How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–39. ISSN: 2573-0142. DOI: 10.1145/3476068.

[414] Sahil Verma and Julia Rubin. “Fairness definitions explained”. In: *Proceedings of the International Workshop on Software Fairness*. New York, NY, USA: ACM, May 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776.

[415] Sandra Wachter and Brent Mittelstadt. “ Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI.” In: *Columbia Business Law Review* 2 (2019), pp. 494–620.

[416] Sandra Wachter, Brent Mittelstadt, and Chris Russell. “Counterfactual explanations without opening the black box: Automated decisions and the GDPR”. In: *Harv. JL & Tech.* 31 (2017), p. 841.

[417] Annie J Wang. “Procedural justice and risk-assessment algorithms”. In: *Available at SSRN 3170136* (2018).

[418] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. “Designing Theory-Driven User-Centric Explainable AI”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–15. ISBN: 9781450359702. URL: https://doi.or g/10.1145/3290605.3300831.

[419] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. "Factors Influencing Perceived Fairness in Algorithmic Decision-Making". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376813.

[420] Zezhong Wang, Jacob Ritchie, Jingtao Zhou, Fanny Chevalier, and Benjamin Bach. "Data Comics for Reporting Controlled User Studies in Human-Computer Interaction". In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (Feb. 2021), pp. 967–977. ISSN: 1077-2626. DOI: 10.1109/TVCG.2020.3030433.

[421] Elizabeth Anne Watkins. "The tension between information justice and security: Perceptions of facial recognition targeting." In: *Joint Proceedings of the ACM IUI 2021 Workshops*. 2021.

[422] Jenny S. Wesche and Andreas Sonderegger. "When computers take the lead: The automation of leadership". In: *Computers in Human Behavior* 101 (Dec. 2019), pp. 197–209. ISSN: 07475632. DOI: 10.1016/j.chb.2019.07.027.

[423] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. "The What-If Tool: Interactive Probing of Machine Learning Models". In: (July 2019). DOI: 10.1109/TVCG.2019.2934619.

[424] Carla Willig. *EBOOK: introducing qualitative research in psychology*. McGraw-hill education (UK), 2013. ISBN: 0335244505.

[425] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 666–677. ISBN: 9781450383097. DOI: 10.1145/3442188.3445928. URL: https://doi.org/10.1145/3442188.3445928.

[426] Langdon Winner. "Do Artifacts Have Politics?" In: *Daedalus* 109.1 (1980), pp. 121–136. ISSN: 00115266. URL: http://www.jstor.org/stable/20024652.

[427] S. Woods, M. Walters, Kheng Lee Koay, and K. Dautenhahn. "Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach". In: *9th IEEE International Workshop on Advanced Motion Control, 2006*. IEEE, 2006, pp. 750–755. ISBN: 0-7803-9511-1. DOI: 10.1109/AMC.2006.1631754.

[428] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang Antony Chen. "CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Vol. 20. 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376807. URL: http://dx.doi.org/10.1145/3313831.3376807.

[429] Pulei Xiong, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. "Towards a Robust and Trustworthy Machine Learning System Development". In: (Jan. 2021).

[430] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. “FairGAN: Fairness-aware Generative Adversarial Networks”. In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 570–575. DOI: 10.1109/BigData.2018.8622525.

[431] Xuhai Xu, Mengjie Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kahyun Kim, and Hrvoje Benko. “XAIR: A Framework of Explainable AI in Augmented Reality”. In: (Mar. 2023). DOI: 10.1145/3544548.3581500.

[432] An Yan and Bill Howe. “Fairness-Aware Demand Prediction for New Mobility”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01 (Apr. 2020), pp. 1079–1087. DOI: 10.1609/aaai.v34i01.5458. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5458.

[433] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. “Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376301.

[434] Vahid Yazdanpanah, Enrico Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, and Timothy Norman. “Responsibility Research for Trustworthy Autonomous Systems”. In: *20th International Conference on Autonomous Agents and Multiagent Systems (03/05/21 - 07/05/21)*. Mar. 2021, pp. 57–62. URL: https://eprints.soton.ac.uk/447511/.

[435] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. “Enhanced Membership Inference Attacks against Machine Learning Models”. In: (Nov. 2021).

[436] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. “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)*. May 2023.

[437] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. “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*. New York, NY, USA: ACM, Apr. 2023, pp. 1–21. ISBN: 9781450394215. DOI: 10.1145/3544548.3581161.

[438] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. “Towards a multi-stakeholder value-based assessment framework for algorithmic systems”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, June 2022, pp. 535–563. ISBN: 9781450393522. DOI: 10.1145/3531146.3533118.

[439] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 335–340. ISBN: 9781450360128. DOI: 10.1145/3278721.3278779. URL: https://doi.org/10.1145/3278721.3278779.

[440] Qiaoning Zhang, Matthew L Lee, and Scott Carter. "You Complete Me: Human-AI Teams and Complementary Expertise". In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022, pp. 1–28. ISBN: 9781450391573. DOI: 10.1145/3491102.3517791.

[441] Angela Zhou, David Madras, Inioluwa Raji Raji, Bogdan Kulynych, Smitha Mili, and Richard Zemel. *Call for participation: Participatory Approaches to Machine Learning*. URL: https://participatoryml.github.io/.

[442] Jianlong Zhou, Sunny Verma, Mudit Mittal, and Fang Chen. "Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making". In: (Sept. 2021).

[443] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. "AI and Ethics – Operationalising Responsible AI". In: (May 2021).

[444] Christian Zimmermann, Rafael Accorsi, and Gunter Muller. "Privacy Dashboards: Reconciling Data-Driven Business Models and Privacy". In: *2014 Ninth International Conference on Availability, Reliability and Security*. IEEE, Sept. 2014, pp. 152–157. DOI: 10.1109/ARES.2014.27.

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

[446] Arkaitz Zubiaga, Bo Wang, Maria Liakata, and Rob Procter. "Political Homophily in Independence Movements: Analyzing and Classifying Social Media Users by National Identity". In: *IEEE Intelligent Systems* 34.6 (2019), pp. 34–42. DOI: 10.1109/MIS.2019.2958393.

[447] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda". In: *Government Information Quarterly* 38.3 (July 2021), p. 101577. ISSN: 0740624X. DOI: 10.1016/j.giq.2021.101577.

Research Material and Data

Research material used as part of this dissertation is publicly available at 4TU.Center for Research Data. Quantitative (anonymized) data generated as part of chapter 4 and chapter 5 (study 2) is also publicly available. We additionally include the living document where the framework presented in chapter 2 is hosted (on Github).

- **Chapter 2**:
  https://github.com/mireiayurrita/valuebasedframework
- **Chapter 3**:
  https://doi.org/10.4121/be171486-fe03-45fe-8d8b-22b4c81cd3a2
- **Chapter 4**:
  https://doi.org/10.4121/62a7ad5f-1225-4618-bd4b-1d66a3941db3
- **Chapter 5**:
  https://doi.org/10.4121/8c19bb03-14de-4c85-b781-33eed0cac44a

Curriculum Vitæ

Mireia YURRITA SEMPERENA

Professional Experience

2025–present	Postdoctoral Researcher, Utrecht University, Utrecht, Netherlands
2021–2024	PhD Candidate, Delft University of Technology, Delft, Netherlands
2020–2021	Hydraulic Design Engineer, Ingeteam, San Sebastian, Spain
2019–2020	Visiting Student, Massachusetts Institute of Technology, Cambridge, USA
2017–2018	Research Intern, Orona, San Sebastian, Spain
2015–2016	Research Assistant, University of Navarra, San Sebastian, Spain

Education

2021–2025	Doctor of Philosophy (PhD), Human-Computer Interaction Delft University of Technology, Delft, Netherlands
2018–2020	Master of Science (MSc), Industrial (mechanical) Engineering University of Navarra, San Sebastian, Spain
2014–2018	Bachelor of Science (BSc), Industrial Technology Engineering University of Navarra, San Sebastian, Spain

Awards and Fellowships

2025	Honorable Mention Recognition at the ACM CHI conference
2023	Best Paper Award at the ACM CHI conference
2023	Best Student Paper Award at the AAAI / ACM AIES conference
2019	*International Connecting Talent* Fellowship – Fomento de San Sebastian
2019	*Global Internship Program* Fellowship – Caja Rural de Navarra
2018	*Universidad de Navarra-Grupo Santander* Fellowship – Banco Santander

List of Publications

Conference and journal publications

9. **Mireia Yurrita**, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, Alessandro Bozzon. “Personalize, Prioritize, Collectivize: Identifying Algorithmic Decision Subjects’ Needs for Meaningful Contestability”. In: *Proceedings of the ACM on Human-Computer Interaction CSCW* (CSCW’25) (2025).

8. Agathe Balayn, **Mireia Yurrita**, Fanny Rancourt, Fabio Casati, Ujwal Gadiraju. “Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use”. In: *CHI Conference on Human Factors in Computing Systems* (CHI’25) (2025). 🏅

7. **Mireia Yurrita**, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia Pont, Alessandro Bozzon. “Towards Effective Human Intervention in Algorithmic Decision-Making: the Effect of Decision-Makers’ Configuration on Decision-Subjects’ Fairness Perceptions”. In: *CHI Conference on Human Factors in Computing Systems* (CHI’25) (2025).

6. Andrea Tocchetti*, Lorenzo Corti*, Agathe Balayn*, **Mireia Yurrita**, Philip Lippmann, Marco Brambilla, Jie Yang. “AI Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities”. In: *ACM Computing Surveys* (2024).

5. Chadha Degachi*, Siddharth Mehrotra*, **Mireia Yurrita**, Evangelos Niforatos, Myrthe Lotte Tielman. “Practising Appropriate Trust in Human-Centred AI Design”. In: *Extended Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* (CHI EA’24) (2024).

4. Kars Alfrink, Ianus Keller, **Mireia Yurrita**, Denis Bulygin, Gerd Kortuem, Neelke Doorn. “Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI”. In: *She Ji: The Journal of Design, Economics, and Innovation* (2024).

3. Agathe Balayn, **Mireia Yurrita**, Jie Yang, Ujwal Gadiraju. “Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developers’ Practices in Handling Algorithmic Harms”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AIES’23) (2023). 🏆

2. **Mireia Yurrita**, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, Alessandro Bozzon. “ Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability”. In: *CHI Conference on Human Factors in Computing Systems* (CHI’23) (2023). 🏆

1. **Mireia Yurrita**, Dave Murray-Rust, Agathe Balayn, Alessandro Bozzon. “Towards a multi-stakeholder value-based assessment framework for algorithmic systems”. In: *ACM Conference on Fairness, Accountability, and Transparency* (FAccT’22) (2022).

Preprints

2. Timothée Schmude, **Mireia Yurrita**, Kars Alfrink, Thomas Le Goff, Tiphaine Viard. “Two Means to an End Goal: Connecting Explainability and Contestability in the Regulation of Public Sector AI”.

1. Miny Rajiv*, **Mireia Yurrita***, Sietze Kuilman, Luciano Cavalcante Siebert, Alessandro Bozzon. “Towards Responsible AI Adoption: Implementation Practices and Challenges in the Dutch Public Sector”.

Workshop papers

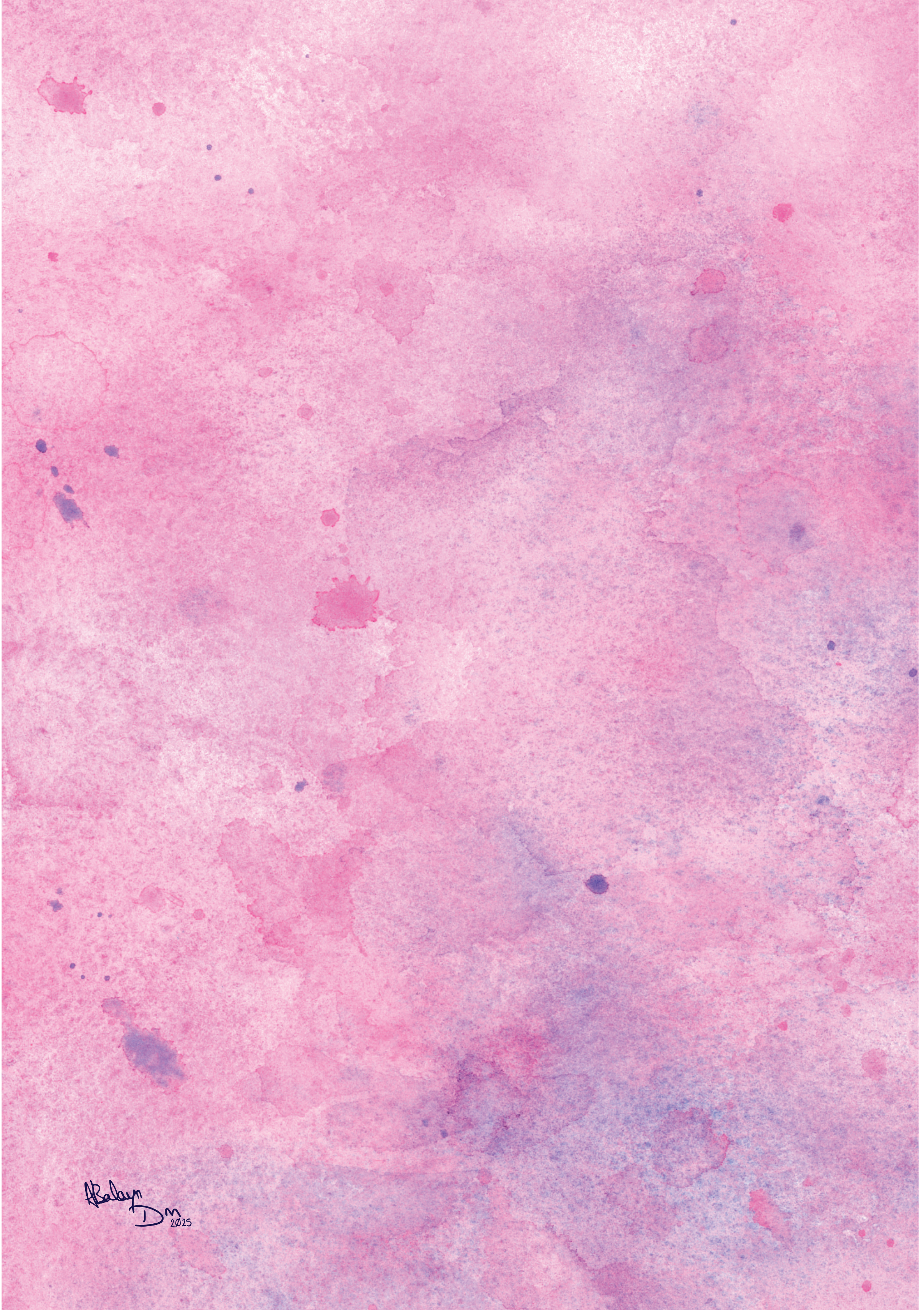
2. Agathe Balayn, **Mireia Yurrita**, Fanny Rancourt, Fabio Casati, Ujwal Gadiraju. “An Empirical Exploration of Trust Dynamics in LLM Supply Chains” In: *Workshop on Trust and Reliance in Evolving Human-AI Workflows at CHI* (TREW’24) (2024).

1. **Mireia Yurrita**, Agathe Balayn, Ujwal Gadiraju. 2023. “Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities”. In: *Human-Centered XAI Workshop at CHI* (HCXAI’23) (2023).

Workshop organization

2. Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, **Mireia Yurrita**, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, Blair Attard-Frost, Ujwal Gadiraju. “From Stem to Stern: Contestability Along AI Value Chains”. In: *ACM Conference on Computer-Supported Cooperative Work and Social Computing* (CSCW’24) (2024).

1. Wesley Hanwen Deng, **Mireia Yurrita**, Mark Díaz, Jina Suh, Nick Judd, Lara Groves, Hong Shen, Motahhare Eslami, Kenneth Holstein. “Responsible Crowdsourcing for Responsible Generative AI: Engaging Crowds in AI Auditing and Evaluation”. In: *AAAI Conference on Human Computation and Crowdsourcing* (HCOMP’24) (2024).


ABelayn DM
2025