The Role of Explanations in Enhancing Algorithmic Fairness Perceptions

Sadaf Afrashteh
(https://orcid.org/0000-0002-0045-6089)

Submitted in total fulfilment for the degree of Doctor of Philosophy

October 2021
School of Computing and Information Systems
The University of Melbourne
ABSTRACT

Decision-makers employ machine learning algorithms for decision-making to gain insights from data and make better decisions. More specifically, advanced algorithms can help organizations classify their customers and predict their behavior at the highest accuracy levels. However, the opaque nature of those algorithms has raised concerns around some potential unintended consequences they might cause, unfair decisions at the center. Unfair decisions negatively impact on both organizations and customers. Customers may lose their trust in organizations as being treated unfairly and consequently, organizations’ reputations might be put at risk.

Transparency provision has been introduced to organizations as a way of addressing the issue of algorithmic opacity. One approach for transparency provision is explaining algorithms’ performance and how they reach a decision to decision-makers. Therefore, explanations can consequently influence the fairness perceptions of the decision-makers about algorithms’ decisions. Understanding how explanations and the way of discoursing them impact on the fairness perceptions of the organizational decision-makers is important. However, little research has focused on the role of explanations in enhancing fairness perceptions.

I seek to address this research gap answering the question of: “How does explanation influence decision-makers’ perceptions of fairness connected with an algorithm’s decisions?” I conduct three studies to answer this question. In study 1, I conduct a conceptual study to explore the dimensions of explanations that need to be studied in understanding the impact of explanations on fairness perceptions. In study 2, I develop a research model hypothesizing the role of perspective-taking in discoursing two different explanations with decision-makers in their fairness perceptions. I conducted a 2*2 experiment to test the hypotheses. In study 3, I develop a research model hypothesizing
the influence of explanations restrictiveness in the decision’s fairness perceived by the
decision-makers. I conducted a 2x2 experiment to test the hypotheses.

The findings of this research result in three important insights about explanations and
their role in enhancing algorithmic fairness perceptions; first, I propose four dimensions
of explanations that need to be considered in understanding fairness perceptions
including the content types of explanations, the explanations reasoning logic, the scope
of explanations and explanations discourse. Second, taking different perspectives of
organization or customer in communicating different types of explanations lead to
different impact on the perception of fairness about algorithm’s performance and its
made decision. Third, Framing explanations in a less restrictive way creates the space
for the decision-makers to be cognitively more engaged with the algorithmic decision-
making and practice their own judgment about that which consequently influences on
their fairness perceptions.
DECLARATION OF AUTHORSHIP

This is to certify that:

i. the thesis comprises only my original work towards the PhD,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.

Sadaf Afrashteh

10 October 2021
PREFACE

This section includes the list of academic articles that I have published and doctoral consortiums that I have attended during my PhD research. Elements of these articles are included in this thesis, particularly Chapter 2 and Chapter 3. The inclusion of the papers are mentioned in relevant chapters of the thesis.

Publications from the PhD Research


Doctoral Consortiuems

ACKNOWLEDGEMENTS

This PhD journey has been a truly life-changing experience for me and it would not have been possible to undertake without the support and guidance that I received from many individuals. Here, it is my pleasure to acknowledge the efforts of all these individuals.

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Michael Davern for the continuous and genuine support of my PhD, for his patience, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. He gave me the opportunity to develop and let me grow as a researcher and for this I owe him a massive gratitude. Thank you for your unwavering support and belief in me.

A special appreciation goes to another supervisor of mine, Dr Ida Someh, whose insight and knowledge steered me through this research. Without her tremendous understanding, treasured support and encouragement in the past few years, it would be impossible for me to complete this research.

I would further take this opportunity to thank Prof. James Bailey, for supervising and providing me with support and guidance in several aspects of my PhD study.

My sincere thank also goes to Prof. Marcello La Rosa for being the chair of my advisory committee and for the insightful guidance that kept me on track with my research. Meanwhile, I would like to acknowledge the University of Melbourne, the School of Computing and Information Systems, for the excellent workplace environment.
Further, I could not have completed this dissertation without the support of my true friends, Vida, Donia, Anam and Francesco, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research. I cannot thank you enough for being there for me through all of the ups and downs of this journey.

Further, I would extend my gratitude to my wonderful husband, Alireza, who instilled in me the virtues of perseverance and relentlessly encouraged me to strive for excellence. You have taught me to work hard for the things that I aspire to achieve. I am truly thankful for having you in my life.

Last but not least, I would also like to thank my amazing family for their continuous support. My most caring mother, Simin, for her unconditional love and unlimited support. My wonderful siblings, Sepideh and Mohammadreza, for their endless love and constant encouragement.
DEDICATION

This dissertation is dedicated to:

my loving and selfless mother who has raised me to be the person who I am today,

the memory of my beloved father, may his memory forever be a comfort and a blessing.
# Table of Contents

Abstract ........................................................................................................................................ iii

PREFACE ...................................................................................................................................... vii

Acknowledgements ..................................................................................................................... ix

Table of Contents ........................................................................................................................ xiii

List of Figures ............................................................................................................................... xvii

List of Tables ................................................................................................................................. xviii

Chapter 1 Introduction .................................................................................................................... 21
  1.1. The Motivation for the Research ......................................................................................... 22
  1.2. The Knowledge Gap, Aim, and Questions Addressed ..................................................... 24
  1.3. The Structure of the Dissertation ..................................................................................... 26
  1.4. Research Contributions ...................................................................................................... 28

Chapter 2 The literature and conceptual development ................................................................. 29
  2.1. Explanations: A Definition and their Importance ............................................................. 30
  2.2. Explanations Discourse .................................................................................................... 31
  2.3. The Literature-Search Strategy ....................................................................................... 32
  2.4. Explanations in Algorithmic Decision-making ............................................................... 33
      2.4.1. Types of Explanation Content .................................................................................. 35
      2.4.2. Scope of Explanations ............................................................................................ 37
      2.4.3. The Reasoning Logic behind Explanations ............................................................. 37
## Table of Contents

2.5. Outcomes of Explanation................................................................. 38
   2.5.1. Fairness of the Algorithm’s Decisions................................. 39
   2.5.2. Operationalization of Fairness ........................................... 40
2.6. Theories Underlying Explanations.............................................. 41
   2.6.1. Discourse Ethics.............................................................. 41
   2.6.2. Perspective-taking ......................................................... 43
   2.6.3. Decisional Guidance vs. Restrictiveness .............................. 43
2.7. The Gaps in Understanding......................................................... 44
2.8. The Position of the Research Project ....................................... 46

### Chapter 3 Enhancing the perceived fairness of algorithmic decision-making via perspective-taking ................................................................. 49

3.1. Orientation to Perspective-taking.............................................. 50
3.2. Theoretical Background ............................................................ 51
   3.2.1. The Development of Perspective-taking ............................... 51
3.3. The Study’s Research Model ...................................................... 53
   3.3.1. Decision Outcomes .......................................................... 53
   3.3.2. Model-Performance Explanations ..................................... 54
   3.3.3. Evidence of the Reasoning ............................................... 56
3.4. The Research Method ................................................................. 58
   3.4.1. Participants ................................................................. 58
   3.4.2. The Task and Procedures in the Experiment ......................... 59
   3.4.3. Materials for the Experiment ......................................... 61
3.5. Data Analysis and Results ......................................................... 64
   3.5.1. Descriptive Data ........................................................... 64
   3.5.2. Manipulation Checks ....................................................... 65
   3.5.3. Correlations ...................................................................... 66
Table of Contents

3.5.4. Analysis of Covariance ................................................................. 68
3.6. Discussion and Conclusions............................................................ 71

Chapter 4 Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations......................................................... 75

4.1. Introduction to the Restrictiveness Study ........................................ 76
4.2. Background on Related Theory ...................................................... 77
  4.2.1. Restrictiveness .............................................................................. 77
  4.2.2. Decisional Guidance .................................................................. 78
  4.2.3. Decisional Guidance or Restrictiveness ...................................... 79
  4.2.4. Interactivity .................................................................................. 79
4.3. The Research Model ......................................................................... 80
  4.3.1. Decision Outcomes ..................................................................... 81
  4.3.2. Interactive Model-Performance Explanations .............................. 82
  4.3.3. Interactive Counterfactual Explanations ...................................... 83
4.4. The Research Method ...................................................................... 85
  4.4.1. Participants .................................................................................. 85
  4.4.2. The Task and Procedures in the Experiment ............................... 86
  4.4.3. Materials for the Experiment ...................................................... 87
4.5. Data Analysis and Results.................................................................. 91
  4.5.1. Descriptive Statistics .................................................................. 91
  4.5.2. Manipulation Checks .................................................................. 93
  4.5.3. Correlations ............................................................................... 93
  4.5.4. ANCOVA Results ...................................................................... 94
4.6. Discussion ......................................................................................... 101

Chapter 5 Conclusions and directions for the future Error! Bookmark not defined.
Table of Contents

5.1. Summary of Research Findings ................................................................. 105
5.2. Contributions and Implications for Scholarship ................................. 108
5.3. Implications for Practice ......................................................................... 109
5.4. The Project’s Limitations and Openings for Future Work ................. 110
LIST OF FIGURES

Figure 1.1: The overall structure of the research ................................................................. 28

Figure 2.1: Explanations Discourse ..................................................................................... 31

Figure 3.1: The model developed for the first empirical study ............................................. 53

Figure 4.1: Proposed Research Model .................................................................................. 81

Figure 5.1: Potential future work on sources of explanations’ restrictiveness ............ 112
LIST OF TABLES

Table 3.1: Participant characteristics ............................................................................. 59
Table 3.2: Material used in the experiment ....................................................................... 61
Table 3.3: Rotated component matrix ................................................................................ 62
Table 3.4. “Faith in General Technology” Factor Analysis Results. ............................... 63
Table 3.5: Descriptive statistics for the dependent variables .......................................... 64
Table 3.6: The decisions’ choice frequencies ................................................................... 65
Table 3.7: Correlation matrix for the variables .................................................................. 67
Table 3.8: ANCOVA results where the dependent variable is perceived reasonability with the system’s performance .............................................................. 69
Table 3.9: ANCOVA results where the dependent variable is perceived satisfaction with the system’s performance ................................................................. 69
Table 3.10: ANCOVA results where the dependent variable is perceived fairness with the system’s performance ................................................................. 70
Table 3.11: ANCOVA results where the dependent variable is the level of perceived reasonability with the algorithm’s decision .................................................. 70
Table 3.12: ANCOVA results where the dependent variable is the level of perceived satisfaction with the algorithm’s decision .................................................. 70
List of Tables

Table 3.13: ANCOVA Results (Dependent Variable: Perception of Fairness about Algorithm’s Decision) ........................................................................................................71

Table 4.1: Example presentation of comparative data

Table 4.2: Participant characteristics .................................................................86

Table 4.3: Material used in the experiment........................................................87

Table 4.4: Rotated component matrix ................................................................89

Table 4.5: Factor-analysis results for general faith in technology ...................90

Table 4.6: Factor-analysis results for need for cognition ..................................90

Table 4.7: Descriptive statistics for the dependent variables ............................92

Table 4.8: The decisions’ choice frequencies ....................................................92

Table 4.9: Correlation matrix for the variables ...............................................94

Table 4.10: ANCOVA results with perceived reasonability of the system’s performance as the dependent variable .................................................................97

Table 4.11: Results of ANCOVA where the dependent variable is perceived satisfaction with the system’s performance .................................................................97

Table 4.12: Results of ANCOVA where the perceived fairness of system performance is the dependent variable .................................................................98

Table 4.13: ANCOVA results where the dependent variable is perceived reasonableness of the algorithm’s decisions .................................................................98

Table 4.14: Results of ANCOVA where satisfaction perceptions related to the algorithm’s decisions are the dependent variable .........................................99

Table 4.15: ANCOVA results with the algorithmic decision’s perceived informational fairness as the dependent variable .........................................................99
List of Tables

Table 4.16: ANCOVA results with the algorithmic decisions’ perceived procedural fairness as the dependent variable

Table 4.17: ANCOVA results where algorithmic-decision-related perceived distributional fairness is the dependent variable

Table 5.1: Summary of the first empirical study’s results

Table 5.2: Summary of the second empirical study’s results
Chapter 1
INTRODUCTION

Businesses make increasing use of algorithms for their decisions and organizational tasks (Benbya, Pachidi, & Jarvenpaa, 2021). For example, entities in the financial sector employ them for fraud detection: the algorithms are able to recognize signs of fraud in pending transactions upon being trained with data that encompass historical patterns (Davenport, 2018). Recommendation systems in retail that suggest further products and services to customers can be cited as another example. These systems afford organizations better customer profiling and personalization (Davenport & Ronanki, 2018). Some more advanced algorithms, such as those applying deep-learning techniques, have reached the highest possible levels of accuracy in their predictions and classification, thanks to their ability to perform complex operations across multiple domains, from health care and finance to education, by utilizing a broad range of data types, including text, images, and sound (Chakraborty et al., 2017). These algorithms’ design is inspired by operations within the human brain. They take a massive quantity of raw data as input and, as the human learning process does, perform a task repeatedly through their multiple layers, making progress each time until reaching the outcome. This is a self-learning process, free of any human intervention, through which the algorithm “tries” to find a structure in the data by detecting useful features. The lack of transparency inherent to such algorithms makes it difficult to understand them, even for data scientists (Montavon, Samek, & Müller, 2018; Ras, van Gerven, & Haselager, 2018). This opaque nature of the algorithms has raised concerns around potential for unintended consequences, unfair decisions among them. Organizations’ provision for transparency of how accurately their algorithms perform and how they reach the specific decision produced has been put forward as the principal way of addressing the opacity issue, with various parties having proposed solid explanations as a central means of providing transparency of algorithmic decision-making for its users. Insights gleaned via the explanations offered influence later decision-makers’ perceptions of the algorithmically made decisions, including the sense of satisfaction and the perceived fairness. Despite this crucial implication of explanations and the ways they need to be communicated, little research has investigated these effects.

The research project presented here explores the impact of explanations on organizational decision-makers’ perceptions as to fairness of the algorithms’ output, where explanations are defined as “the meaningful information about the logic involved” (Goodman & Flaxman, 2016, p. 6). The work behind the dissertation focused
on examining what constitutes meaningful information and how it needs to be framed if it is to make an impact on decision-makers’ sense of how fair the relevant decision is. The sections below articulate the importance of this topic, outline the focus of the research undertaken to address it, describe the project’s structure, and introduce the key contributions made.

1.1. The Motivation for the Research

Addressing explainability of opaque algorithms is important for three main reasons. Firstly, growing use of algorithms by organizations in many spheres of life has raised concerns about their implications for the individuals affected, society and the organizations themselves (Beer, 2009; Dressel & Farid, 2018; Greenfield, 2010; Kitchin, 2017; Zarsky, 2015), implications that present ethics issues related to matters such as unfairness (Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016), traceability (Mittelstadt et al., 2016), accountability (Lepri et al., 2018; Shin & Park, 2019), and surveillance facilitated by algorithms (Zuboff, 2015). Experience has already produced several examples of unfairly discriminatory decisions that disfavor a disadvantaged or vulnerable group of individuals (Barocas & Selbst, 2016). One of these is the biased prediction of recidivism by COMPAS, a system designed to predict the risk that a criminal will re-offend. Per COMPAS predictions, blacks are almost twice as likely as whites to re-offend (Dressel & Farid, 2018). Unfair decisions bring disadvantages both for those subject to the decisions (such as customers), including distrust, and for the organizations that use the algorithms, among them reputation risks (Afrashteh, Someh, & Davern, 2020). The algorithms’ fairness may be called into question whenever there is little or no understanding of how they work. Therefore, explanations play a significant role here, addressing the issue of opacity by shedding light on how algorithms reach their decision.

Secondly, the European Union in its General Data Protection Regulation (GDPR) has granted individuals a right to explanation, characterized as “meaningful information about the logic involved in the data analysis” (Goodman & Flaxman, 2016, p. 6), as a means of increasing transparency for the people whose data have been used. Therefore, organizations that utilize personal data of their customers in decision-making are obliged to adhere to the associated regulations. For instance, customers of a bank requesting for a loan approval are allowed to ask for transparency about how the loan approval or denial decisions have been made when it is necessary. However: there is not
much clarity of what exactly this meaningful information is and how its content has to be explained. Along with the regulations introduced in European Union, Australian government, the department of industry, science, energy and resources has also introduced Australia’s AI ethics principles consisting of human, societal and environmental wellbeing, human-centred values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability and accountability. These principles are the main pillars of the australia’s ethical AI voluntary framework which are developed to ensure AI is safe, secure and reliable("Australia's AI Ethics Framework," 2018).

The third factor motivating this research is a desire to respond to information-systems (IS) scholars’ call for contributing to the explainability of algorithmic decision-making. For nearly three decades, IS researchers have been studying explanations, across a range of intelligent systems that covers decision-support systems, expert systems, and recommendation systems (Arnold, Clark, Collier, Leech, & Sutton, 2006; Berendt & Preibusch, 2014; Gregor & Benbasat, 1999); however, domain experts have defined and set the rules for these systems, and with the emergence of machine learning they learn independently and the rules are generated from data. Recently, several special issues of top journals and editorials from well-known researchers have dealt with explanations for algorithmic decision-making and the research opportunities available to IS researchers. These reinforce the call for research focusing on ethics and addressing the issues around biased and unfair decisions, accountability, and explainability of algorithms (Benbya, Davenport, & Pachidi, 2020; Benbya et al., 2021).

Notwithstanding scholars’ strong interest in explanations connected with algorithmic decision-making, there is little understanding of how explanations influence the organizational decision-makers’ perceptions of fairness. Furthermore, outside academia, practitioners are keen to understand what kind of information facilitates fuller understanding of the decisions made by algorithms, an improvement that leads to change in the decisions’ perceived fairness. Therefore, I carried out research investigating the impact of algorithmic decisions’ explanations on decision-maker perceptions of fairness.
1.2. The Knowledge Gap, Aim, and Questions Addressed

In response to concerns related to algorithms’ explainability and transparency, a significant amount of research has been devoted to how to address these issues (Ananny & Crawford, 2018; Arnold et al., 2006; Christensen & Cheney, 2015; Ghani, 2016; Gilpin et al., 2018; Rader, Cotter, & Cho, 2018; Shin & Park, 2019; Wachter, Mittelstadt, & Russell, 2017). Current literature can be divided into two streams, one centered on explainable artificial intelligence (XAI) and the other on the explaining the AI-enabled decisions. The XAI literature focuses on developing machine-learning techniques to apply more explainable and interpretable algorithmic models while maintaining the highest possible level of accuracy (Adadi & Berrada, 2018), whereas literature in the second class considers the explanations and their impact on the behavior of the algorithms’ users and subjects, among them organizations’ decision-makers and customers. Individuals’ tendency to make use of AI-enabled decisions can be impacted by their perceptions about the AI outcomes (Benbya et al., 2021). Such studies investigate distinct types of explanations too, together with their presentation format and provision mechanisms (Giboney, Brown, Lowry, & Nunamaker, 2015; Gregor & Benbasat, 1999; Li & Gregor, 2011; Martens & Provost, 2013). Also, several pieces of research have studied explanations as a mechanism of enhancing transparency to address accountability issues (Doshi-Velez et al., 2017; Mittelstadt, 2016; Mittelstadt et al., 2016). However, the link between the explanations and the perceptions of fairness has not been explored.

Much of the prior literature examines various technical mechanisms designed to enhance algorithms’ fairness (Bakalar et al., 2021; Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018). The authors mainly argue that the solution is to filter out sensitive attributes such as gender and race from the dataset via which the decision is derived (Barocas & Selbst, 2016; Schermer, 2011). This does not always help, though. For example, in some cases, correlations between these attributes and the remaining ones can “leak” connections, thereby leaving the dataset with implicit bias.

An alternative is to mitigate the risks of unfair decisions by maximizing transparency. The term “transparency” refers specifically to the extent to which the algorithm is open in communicating the data used and the logic employed to reach a decision (Lepr
In my work, I argue that the explanations provided as a means of transparency influence the algorithm-users’ perceptions of fairness.

The aim of this research was to reveal how one can influence an organization’s decision-makers’ perceptions as to the fairness of the decisions by using explanations. This is important because once the explanations can shed light on the algorithmic decisions and the perceived fairness is influenced by them, the organizational decision-makers can take corrective actions in case they perceive the decision as unfair. Therefore, the main research question was posed as follows:

*How does explanation influence decision-makers’ perceptions of fairness connected with an algorithm’s decisions?*

Answering this question requires answers to three sub-questions, explained next.

For addressing the main research question, it is important, firstly, to understand what constitutes the meaningful information for purposes of the explanations. Diverse examples of the information considered to be required for explanation can be found in prior literature (Arnold et al., 2006; Gregor, 2001; Gregor & Benbasat, 1999; Sørmo, Cassens, & Aamodt, 2005; Wachter et al., 2017; Wang & Benbasat, 2007). This is contextual, however: which information has meaning varies with the level of expertise of the explanation-receiver and with the objective for explanations’ provision. Different types of explanations can guide a decision-maker toward different goals (Benbya et al., 2021; Silver, 1990). For example, one study shows that the explanation required by novices if they are to understand a knowledge-based system differs from that required by expert users, and the two differ in their impact on human decision-making behavior (Arnold et al., 2006). Despite several studies exploring the various possible types of explanation content, though, scholars still have produced little clarity around what meaningful information needs to be communicated to an organization’s decision-makers about algorithms. Therefore, the first sub-question tackled in the research is this:

- **What is the meaningful information that must be explained to the decision-makers at an organization when it employs algorithmic decisions?**

Once the required meaningful information has been clarified, there remains a need to explore how the decision-makers’ perceptions of fairness can be influenced by explanations. One approach is to frame the explanations through perspective-taking, a
cognitive process that can be defined as considering the attitudes and views of someone else within a given situation and trying to comprehend that situation through seeing his or her needs and beliefs (Lee, Lee, & Keil, 2018; Te‘eni, Sagie, Schwartz, Zaidman, & Amichai-Hamburger, 2001). Importantly, taking someone else’s perspective can result in behavior change that benefits the target of the perspective-taking (Galinsky, Wang, & Ku, 2008). Among the stakeholders that might be engaged as targets in the process of decision-making are the organization and customer. Since taking different stakeholders’ perspective in the communication of the explanations results in multifaceted perceptions of the decisions’ fairness, the second sub-question for this project can be expressed thus:

- How does perspective-taking in explaining an algorithm’s decisions influence the organizational decision-makers’ perceptions of fairness?

While explanations are introduced to enhance transparency, there has always been tension between strong guidance for the decision-makers, with their decision space being restricted against the background of explanations, and, on the other hand, creating a space for them to practice their own decision-making within a framework of less overtly guiding explanations (Benbya et al., 2021; Silver, 1990). Therefore, for those aiming to influence perceptions as to the fairness of the algorithm’s decision, it is important to know the extent to which the explanation needs to be restrictive and explicitly guiding vs. create room for the decision-makers to exercise their judgment:

- How does restrictive explanation influence the fairness perceived by organizational decision-makers?

The three subsidiary questions specified above served as signposts on the path of clarifying the impact of explanations on organizational decision-makers’ perceptions of the fairness of algorithmic decisions and, thereby, addressing the main research question.

1.3. The Structure of the Dissertation

This dissertation comprises five chapters. The next one provides a detailed introduction to the relevant background for the research, including a literature review, and develops the conceptual grounding. Then, chapters 3 and 4 present the empirical studies conducted for understanding the impact of explanations on the decision-makers’ sense
of algorithmic decisions’ fairness: Chapter 3 examines how communicating the explanations on the basis of stakeholder perspectives influences decision-maker perceptions of fairness, by describing an experiment in this area, and Chapter 4 presents an experimental study in which I investigated the impact of less restrictive explanations on perceived fairness. Finally, Chapter 5 explores the key contributions, limitations, and potential directions for further work connected with this research.

A brief outline of the role of each chapter may help to orient the reader.

Providing a solid introduction to the relevant background for this research and the literature informing it, Chapter 2 sheds light on the conceptual development of the research. It provides a starting point by supplying background on explanations for algorithmic decision-making and the underlying theories pertaining to explanations.

Chapter 3 presents the first experiment, focusing on the impact of explanations framed through perspective-taking on the decision-makers’ perceptions of fairness. This chapter reveals the importance of perspective-taking in the explanations communication. Along with the relevant theoretical background, it details the experiment design and data-collection work that led to the conclusions drawn from the experiment’s results.

The second experiment is presented in Chapter 4. That empirical work focused on the importance of the tension visible in current practice that pits largely guidance-oriented explanations against less restrictive ones. The chapter examines the impact of both on a decision-maker’s perceptions of how fair the algorithm’s decision is. This chapter too provides both the relevant theoretical background information and a detailed description of the experiment’s design, collection of data, and results and findings.

Chapter 5 provides integrative discussion of the key findings from the research. The chapter also functions as a conclusion for the dissertation, presenting the answers to the research questions, characterizing the overall contribution of the research, and outlining both the limitations associated with this research and recommendations for future work.
1.4. Research Contributions

The research project contributes to scholarship addressing algorithmic decisions’ explainability, with particular focus on current concerns about unfair decisions made by algorithms – decisions whose unfairness could get obscured by their opaque nature.

The key contributions represented by this dissertation can be summarized as primarily lying in better understanding of how explanations can influence the perceptions of organizations’ decision-makers as to the decisions’ fairness. The studies made strides in this direction by exploring what type of information is suited to making an impact on the sense of fairness. The dissertation also introduces two important strategies that need to be considered in efforts at communicating the explanations well: perspective-taking and explanations’ restrictiveness. Both influence the fairness that decision-makers perceive in the algorithm’s decisions and create space for these human actors to practice discretion and make more informed decisions of their own.
Chapter 2
THE LITERATURE AND CONCEPTUAL DEVELOPMENT

This chapter\(^1\) sets the scene by considering what explanations are and why they are needed. It then presents a synthesis of preexisting literature on explanations, within the domain of IS research. This integrative discussion aids in identifying the dimensions associated with explanations that one must consider when examining explanations’ impact on the fairness that an organization’s human decision-makers associate with the decision made. It also discusses the issue of fairness itself, reviewing this in connection with algorithms’ decisions and characterizing the relevant studies. Next, key theories underlying the presentation of explanations are outlined, for further theoretical grounding of the study. In light of prior work, multiple research gaps related to explanations and their influence on the perceived fairness of algorithm-produced decisions are identified, in the final part of the chapter. This orientation to the research landscape lays the groundwork for later chapters.

\(^1\) Elements of the chapter have been published in the following article:
2.1. Explanations: A Definition and their Importance

Explanations, in general, can be described as means of clarification and improvement in understanding of a phenomenon (Gregor & Benbasat, 1999). Under one definition, an explanation is a “declaration of the meaning of words spoken, actions, motives, etc., with a view to adjusting a misunderstanding or reconciling differences” (Dictionary, 1981, p. 628). A facility for such declarations within the expert-systems context increases the users’ understanding of the system and its decision-making process, thereby improving the usefulness of the expert systems and contributing to user confidence in the decisions produced.

Providing explanations to system-users proves especially important where they could inform decisions or problem-solving, when an anomaly is perceived, and when learning is required (Gregor & Benbasat, 1999). In the context of algorithmic decision-making, explanations can support understanding of how and why a decision has been made, whether by “white-box” algorithms or by more advanced but opaque “black-box” ones. Greater understanding on individuals’ part can enhance acceptance of the decisions made by algorithm-based systems and help human decision-makers improve the algorithms’ performance in the event that errors are found (Martens & Provost, 2013).

Explanations have seen extensive use by IS researchers in the context of expert systems. However, the nature of IS artifacts is changing, with evolution from rule-based systems to application of algorithms that generate the rules from data rather than obtaining them from domain experts (Martens & Provost, 2013). Advanced algorithms that employ artificial intelligence (AI), such as deep-learning tools, function as black boxes in that they learn for themselves by applying unknown inputs to recognize hidden patterns in the data. This makes AI-based explanations a new breed in several respects – e.g., with regard to content and reasoning logic (Ras et al., 2018). This is why the concept of explanation requires more thorough understanding for today’s context. For this aim, a comprehensive literature review was conducted. Its design and search strategy are described below.
2.2. Explanations Discourse

To answer the first research sub-question of “what is the meaningful information that needs to be explained to organizational decision-makers?”, I propose that it is important to consider the explanations discourse. The reason is that in influencing the decision-makers’ perceptions about algorithms’ decision only considering the information which are forming the explanations is not sufficient. The way this information required to be communicated is also of significance. The explanations discourse consists of four dimensions as “content of explanations”, “reasoning logic of explanations”, “scope of explanations”, and “explanations communication strategies”. Figure 2.1 presents the dimensions of explanations discourse.

![Figure 2.1: Explanations Discourse](image)

Content of explanations refers to the type of information provided by explanations. Depending on which aspect of the algorithm needs to be explained, this information varies (Gregor & Benbasat, 1999; Guidotti et al., 2018; M. Silver, 2006; Sørmo, Cassens, & Aamodt, 2005). Scope of explanations refers to the extent of the algorithmic decision-making area that needs to be explained (Fawcett & Provost, 1997; Martens & Provost, 2013). Reasoning logic of explanations refers to the logic employed by explanations to convey the information (Hossin & Sulaiman, 2015; Miller, 2018; Wachter, Mittelstadt, & Russell, 2017) and the communication strategies refers to strategies taken in communicating the explanations to achieve the goal of communication (Te'eni et al., 2001).

Figure 2.1 illustrates that in creating meaningful information for explanations it is important to understand which aspects of algorithm within which scope needs to be
explained and what would be the appropriate logic. Eventually, as the aim of this study is to influence the organizational decision-makers’ perception of fairness, it is important to have a strategy in communicating explanations. All four dimensions are interdependent; depending on which strategy is taken for explanation communication, the content, the reasoning logic and the scope of explanations can be different through the discourse.

2.3. The Literature-Search Strategy

Adopting a systematic approach, I conducted an exploratory literature search to develop a comprehensive literature review addressing explanations. That entailed searching for key papers in the IS basket of eight journals (from the Association for Information Systems) and the Scopus database. This search space covered the top journals and most widely accepted venues in the field that publish high-quality articles. I began by searching for “explanation” or “explainability” in the full text of Management Information Systems Quarterly (MISQ) and Journal of the Association for Information Systems (JAIS) articles and found 107 papers. I reviewed all the papers for their use of the concept of explanation, then removed those papers that limited the discussion of explanation to a general sense; 25 papers remained for further analysis. Then, I expanded the search, looking for “explanation” in the title, abstract, and keywords from Journal of Information Technology (JIT), Information Systems Research (ISR), Journal of Management Information Systems (JMIS), Journal of Strategic Information Systems (JSIS), European Journal of Information Systems (ECIS), and Information Systems Journal (ISJ) papers. This search yielded 58 papers. I filtered out those papers not related to the decision-making context and was left with 17 papers for further analysis.

As for the Scopus material, I searched the database – again, looking for articles with “explanation” or “explainability” in the title, abstract or keywords – to augment the corpus with more top-quality academic publications. The search was limited to the area of business, management, and accounting. Exclusion of papers in which “explanation” and “explainability” served only as general terms resulted in 132 papers. Narrowing the focus to mainstream work reflecting a modern AI context, I removed those published before 2007 or with fewer than 25 citations. Eleven papers were retained for further analysis.
For greater coverage, I utilized Google Scholar too in the exploratory search. I searched for work addressing explanation in combination with AI, algorithms, machine learning, or algorithmic decision-making, then carried out another search iteration, for publications citing that work. For more coverage in the other direction, I also looked at the papers cited in the articles initially returned. All duplicate articles were filtered out.

My review encompassed MISQ, MISQE, and JAIS special issues on AI. Therefore, the project included the two associated editorials from these (Benbya, Davenport, & Pachidi, 2020; Benbya et al., 2021) and entailed reviewing the papers submitted to the special issues. This literature constitutes a solid grounding in the studies that have contributed both practically and theoretically to IS-related AI literature. Of the papers published in these special issues, my review included those that focus on consequences of employing AI and on algorithms’ explainability (Asatiani et al., 2021; Mayer, Strich, & Fiedler, 2020).

2.4. Explanations in Algorithmic Decision-making

Advanced algorithms often operate as black boxes. In the context of Big Data analytics, this is often because the machine-learning algorithm’s way of learning is shrouded from human view, or inscrutable; it follows an inherently opaque decision process (Sinha & Swearingen, 2002). Even when a human-interpretable model and algorithm exist, organizations’ policy may not permit disclosing the workings of the algorithm to internal users, let alone the subjects of the data (e.g., for competition reasons). Whether inherent or dictated by policy, opacity of algorithmic decision-making readily leads users and data subjects alike to distrust the algorithmic model and its decision (Ribeiro, Singh, & Guestrin, 2016). More importantly, to the extent that said decision-making is opaque, it is hard to assure of full, appropriate accountability for the decisions made. This has led to calls for greater transparency of algorithmic decision-making. While transparency indeed enables fuller accountability for the decisions made, thereby enhancing trust (Lepri et al., 2018), it can be problematic, particularly when the machine-learning methods employed are opaque by nature.

One way of addressing the transparency issue is to provide explanations. Systems’ ability to provide explanations of system decisions or recommendations has long been
critical for user acceptance of the decisions and recommendations produced (Dhaliwal & Benbasat, 1996; Hayes-Roth & Jacobstein, 1994; Ye & Johnson, 1995). From the outset, explanations have been sought, where these are defined as providing useful descriptions of the need for the data requested, the reasoning employed in processing of the data, and the basis for any recommendation or decision (Clancey, 1983). Explanations have been found to help users in the three distinct ways alluded to above: 1) catering for users’ needs when their expectations are not fulfilled, 2) facilitating learning, and 3) delivering information required for problem-solving and decision-making (Gregor & Benbasat, 1999).

Explanations have been studied across multiple disciplines, with computer science (CS) and IS being among the most prominent (Adadi & Berrada, 2018; Gregor, 2001; Gregor & Benbasat, 1999). The body of work by computer scientists into XAI to address the explainability issue related to AI algorithms is expanding especially rapidly. In response to the issue, the Defense Advanced Research Projects Agency (DARPA) launched its XAI program in May 2017, for designing new AI models that the machine-learning algorithms can explain themselves, consequently affording understanding by their end users (Prabhakar, 2017). The models can be explained in terms of the overall rationale behind their decision-making or via presentation of the weaknesses and strengths of their performance. While such work holds promise, my study followed a path more aligned with IS literature, so as to build on work that takes a socio-technical approach to explanations, in keeping with that field’s traditions, rather than follow the XAI literature (Gregor & Benbasat, 1999) in its emphasis on investigating machine-learning techniques and other means of developing explainable AI models. The core reason for this decision is that I wished to study the interface with humans – the impact of giving explanations on decision-makers’ perceptions of the decisions made by algorithms.

A vast range of intelligent information systems employs explanations, with some examples being knowledge-based systems (Arnold et al., 2006; Dhaliwal & Benbasat, 1996; Gregor, 2001; Gregor & Benbasat, 1999; Mao & Benbasat, 2000; Smedley & Sutton, 2007), decision-support systems (Berendt & Preibusch, 2014; Gönül, Önal, & Lawrence, 2006; Tan, Tan, & Teo, 2012), expert systems (Arnold et al., 2006; Bohanec, Kljajić Borštnar, & Robnik-Šikonja, 2017; Ye & Johnson, 1995), recommendation agents (McSherry, 2005; Sinha & Swearingen, 2002; Wang & Benbasat, 2007), and case-based reasoning systems (Sørmo et al., 2005). In all these cases, both machines and
individual human agents play a part in the decision-making and problem-solving process. Explanations assist human decision-makers by reducing their cognitive load in the decision-making process and enhancing the quality of the decisions and recommendations made (Gregor, 2001). From the users’ perspective, explanations contribute to more positive perceptions of the system and can also assist with learning in the problem domain (Gregor & Benbasat, 1999), primarily by enhancing transparency. In the consumer-facing context of recommendation agents, this transparency aids in building customer trust in the systems’ recommendations by exposing the logic of the underlying processes to consumers (Sinha & Swearingen, 2002).

Another key issue addressed in prior literature is the role of the user’s expertise in the application domain. The effects of explanations vary significantly, depending on the expertise of the user. Feed-forward explanations conveying declarative knowledge are more useful for novices, satisfying a desire for information on the system’s inputs and their relationships, while experts prefer feedback explanations, which facilitate transfer of procedural knowledge. Explanations of the latter type provide information on what lies between the inputs and outputs: the process by which the given decision has been made (Arnold et al., 2006).

Since my main goal was to explore the impact of explanations on the perceived fairness of algorithms’ decisions, I had to gain a comprehensive understanding of explanation and the various aspects of it. These aspects include the content of the explanation and the information it could convey (Gregor & Benbasat, 1999; Guidotti et al., 2018), the scope covered by the provision of information (Adadi & Berrada, 2018; Martens & Provost, 2013), and the logic followed in the reasoning for the explanations (Hossin & Sulaiman, 2015; Wachter et al., 2017). The following sections cover these aspects, together with the potential outcomes of explanations.

### 2.4.1. Types of Explanation Content

There are several approaches to categorizing explanations in terms of their content. Various taxonomies enable discussing explanations with regard to what information they provide and how this information could shed light on the decision-making process and its outcome (Gregor & Benbasat, 1999). One approach, employed by Gregor and...
Chapter 2 The literature and conceptual development

Benbasat (1999), is to group explanations’ content into four categories of information supplied to users: terminology, trace explanations, justification, and strategy. Terminological explanations convey “knowledge of the concepts and relationships of a domain that experts use to communicate with one another” (Swartout & Smoliar, 1987). Trace explanations describe, ex post, the actual reasoning process employed by the system to reach a specific decision or conclusion. Justification-oriented explanations, in turn, articulate the basis for the conclusion or recommendation made by the system. This is often anchored in domain-specific knowledge. Finally, strategy explanations explain the goals for the system that underlie its reasoning process (Sørmo et al., 2005).

A taxonomy parallel to that of Gregor and Benbasat is found in the literature on decisional guidance, the body of work dealing with how a system enlightens or persuades users in their decision-making (Silver, 1991). Its key proponent, Silver, distinguishes between the concept of informative guidance, designed to enlighten users, and suggestive guidance, intended to sway them (Silver, 2006).

Alternatively, explanations can be categorized by the problems that may be addressed via the information they provide. Here, three types of explanation can be considered: model explanations, describing the entirety of the logic behind the classifier; outcome explanations, which address the reasons for the decisions made on a given object; and model-inspection explanations, which are of use when the goal is to understand how a black box behaves internally when particular changes are made to the system’s input (Guidotti et al., 2018).

Alongside the taxonomies of explanations’ content, it is important to consider how much information needs to be provided. Decision-makers can experience information overload: when excessive information is provided, it may be difficult to assimilate the information, and less effective decisions may result (Gross, 1964). The reason behind the decreased decision quality is that the information input exceeds the human’s cognitive processing capacity. In light of the information-overload theory, it is vital to consider how much information is being communicated, if wishing the explanations to have the greatest impact on decision-makers’ perceptions of algorithmic decisions.
2.4.2. Scope of Explanations

Explanations vary in scope, depending on whether the recipient must understand the entire model (global scope) or, instead, just a single prediction requires an explanation (local scope). Providing information about the complete algorithmic model and its performance, global explanations can apply to all instances rather than merely a single one. The strength of local explanations (such as those provided by LIME; see Ribeiro et al., 2016), on the other hand, lies in focusing on a particular decision that has been made for an individual instance on the basis of an individual’s data. This type of explanation can be useful when personalized information about a specific decision is needed. The scope required depends on the application for the explanations. The context of credit decisions serves as a good example: For regulatory compliance, the performance of the algorithms as a whole has to be explained and justified. Simultaneously, if a specific customer seeks information about why and how a specific decision has been made, for purposes of understanding whether it was made properly, it is necessary to provide local explanations (Martens & Provost, 2013). Explaining the algorithms used for fraud detection is another example of settings wherein local explanations are required (Fawcett & Provost, 1997).

Assessing the impact of explanations on perceived fairness requires one to consider the effect of both local and global explanations. The reason is that a user’s sense of how fair a decision by an algorithm is might be influenced by either: the performance of the algorithm as a coherent model or the circumstances of the specific case as a single instance. Therefore, for this research, I worked with both local and global explanations.

2.4.3. The Reasoning Logic behind Explanations

Several options are available for the reasoning behind the information given in explanations. These logics are explored in the “human-like explanations” literature, whose authors emphasize that the logic employed needs to be consistent with the users’ expectations as to the reasoning behind such explanations (Boonzaier, McClure, & Sutton, 2005; Chin-Parker & Cantelon, 2017; Miller, 2019). According to this stream of literature, users prefer explanations that apply a contrastive reasoning logic, because
they want to know why another event has not taken place rather than the one that did occur (Miller, 2019). This is referred to as counterfactual reasoning. Pinpointing the crux of decisions, it can shed light on cases in which a small change in variables leads to significant changes in the resulting decision. This reasoning attunes people to “the close possible worlds” (Wachter et al., 2017). Conversely, factual reasoning describes the most important factors that contributed to the actual decision. Both factual and counterfactual reasoning are considered local explanations, because they incorporate instance-level information on a specific decision. Factual reasoning points out the root features of the decision, while counterfactuals describe the decision boundaries to show which class the decision falls within (Adadi & Berrada, 2018; Hossin & Sulaiman, 2015; Wachter et al., 2017).

The reasoning system behind explanations serves to inform and persuade users (Silver, 2006). In the literature exploring the persuasiveness of explanations, Toulmin’s model of argumentation has been the primary analysis tool (Ye & Johnson, 1995). This model identifies six elements of argumentation: claims, data, warrants, backing, qualifiers, and possible rebuttals. A claim is a state proposed for acceptance, data form the basis for the argument, and warrants provide the connection between the two that justifies the claim. Backing is a mechanism supporting the trustworthiness of warrants in case their validity is doubted. To address issues of the degree of certainty of a claim, qualifiers are employed. The final element of the model, possible rebuttals indicate conditions in which the warrant is not applicable and, hence, the proposed conclusion may be overruled. This model provides a useful framework for considering the components of an explanation – whether it provides data, backing, qualifiers, etc. (Gregor & Benbasat, 1999). It also can aid in identifying the extent to which an explanation can justify an algorithm’s decision by considering which elements of a persuading argument it introduces.

2.5. Outcomes of Explanation

Providing explanations to the users of algorithms – organizations’ decision-makers, customers affected by their decisions, and other stakeholders – provides transparency that can produce outcomes including improved accuracy, accountability, and perceived fairness (Lepri et al., 2018; Rader et al., 2018; Wang & Benbasat, 2007). Firstly,
explanations can illuminate potential sources of error in algorithms’ performance, thus arming the decision-makers to provide feedback on how system performance could be improved (Martens & Provost, 2013). Explanations can at the same time facilitate the accountability of algorithms by shedding light on what data are used and how a decision is reached (Lepri et al., 2018). Together, improved accuracy and greater accountability can lead to a greater sense that the algorithm’s decisions are fair (Lepri et al., 2018). Since the dissertation project involved studying the influence of explanation and various communication strategies of it on perceptions of fairness, the following section examines this concept in detail.

2.5.1. Fairness of the Algorithm’s Decisions

Fairness is the feeling that people experience when they receive their deserved share of the benefits available (Welsh, 2004). In the setting of algorithmic decision-making, the word refers to “lack of discrimination or bias in the decisions” (Lepri et al., 2018, p. 615). There has been considerable debate around issues of fairness in algorithmic decision-making, caused by the complexity of the algorithms’ performance. This debate spans the disciplines of IS, CS, and law (Jarrahi, 2018; Robert, Pierce, Marquis, Kim, & Alahmad, 2020; Veale, Van Kleek, & Binns, 2018; Vellido, 2019; Wang & Siau, 2018), with much of the literature on algorithmic decision-making’s fairness built on two opposite pillars. Some posit that employing algorithms enhances fairness by not involving humans in the process of decision-making, arguing that it yields benefits by eliminating introduction of human judgment and potential bias (Zarsky, 2015). Other scholars, in contrast, voice concerns about the fairness of the outcomes when algorithms are used (Afrashteh, Someh, & Davern, 2020; Lepri et al., 2018; Martin, 2015; Mittelstadt et al., 2016).

Algorithmic decision-making works by building a model from historical training data, a model that can then be applied to future decisions, whether through predictions, classifications, or other action. This process brings with it several factors that can be cited as reasons for the resulting decisions being unfair, such as incorrect labels, sampling bias, and incomplete datasets. The first of these is related to differences between labels applied to attributes in the training data and the dataset to which the
model is to be applied. To the extent that the attribute labels and definitions from the training data are incorrect or diverge from the context of application, the validity of the model is called into question. Individuals may be misclassified or otherwise dealt with inappropriately. In a similar vein, a biased set of sample data can lead to a model that is unfairly discriminatory. For example, if the historical data used for modeling of suburban crime rates comprise a disproportionate number of suburbs with a high crime rate, the model may exhibit bias and suggest inappropriate actions or interpretations when applied for settings not adequately represented in the training data (i.e., suburbs with lower crime rates). Finally, incomplete training, in which important characteristics of the actual population are, at best, only partially present can yield a model with questionable validity. Somewhat ironically, this situation can arise because of concerns related to individuals’ data privacy: processing the data to address ethics concerns may lead to ethically questionable algorithmic decision-making.

2.5.2. Operationalization of Fairness

There are several ways of operationalizing and measuring fairness. Some researchers, regarding fairness as an overarching construct, have used a single scale to measure it (Ambrose & Schminke, 2009; Choi, 2008; Zapata-Phelan, Colquitt, Scott, & Livingston, 2009). An example measurement item for assessing overall fairness is “How fairly treated am I?” (Lind, 2001, p. 223). On the other hand, several studies have broken fairness down to address four dimensions – informational, interactional, procedural, and distributional fairness – with instruments including separate questions to measure each of these dimensions (Bobocel, 2013; Holtz & Harold, 2009; Jones & Martens, 2009). Procedural fairness is related to the logic and intent behind the process through which a decision is made, and informational fairness has to do with whether the information provided in relation to the decision made addresses the requirements of the individuals for whom it was made. Interactional fairness pertains to how respectfully an individual affected by the decision is treated, while distributional fairness involves a broader perspective: the extent to which the decisions are made equitably for the individuals in view of their specific conditions and situations (Binns et al., 2018).

Another relevant body of work is focused on perceptions of fairness in humans’ decision-making (Binns et al., 2018; Holstein, Wortman Vaughan, Daumé, Dudik, &
Chapter 2 The literature and conceptual development

Wallach, 2019; Lind & Tyler, 1988; Potter, 2006; Rahwan et al., 2019). There is an important distinction between fairness and perceived fairness, in that a decision might be objectively fair but viewed as unfair by an individual for reason of factors such as the negative impacts the decision may have personally (Binns et al., 2018). Because the dissertation project dealt with explanations’ impact on decision-makers’ perceptions of fairness, actual fairness was only a peripheral consideration. This is because the algorithmic decisions remained constant. The explanations could not affect their fairness since they could not change the algorithmic decision-making process. What they could influence is the decision-makers’ perceptions, on the basis of their information content and presentation.

2.6. Theories Underlying Explanations

Within the IS literature, several theories get employed to support the theoretical grounding of research on explanations. Among them are Toulmin’s model of argumentation, the “adaptive control of thought” rational theory, cognitive-effort theory, and theory of interpersonal communication (Arnold et al., 2006; Gregor, 2001; Mao & Benbasat, 2000; Wang & Benbasat, 2007). Most of these theories are tools in studying the impact of explanations’ content and structure on cognition and acceptance among users. For example, scholars have applied cognitive-effort theory to investigate the aforementioned relationship between the expertise of knowledge-based systems’ users and the type of explanation they expect to receive (Mao & Benbasat, 2000). Since the aim of my work was to examine how explanations influence perceptions of fairness, I selected alternative bodies of theory instead. My proposal that explanations can influence perceived fairness draws on discourse ethics. I also bring in perspective-taking theory, to explore how explanations can influence perceptions of fairness. Finally, decisional guidance theory aids in clarifying whether explanations’ restrictiveness plays a role in the degree of perceived fairness.

2.6.1. Discourse Ethics

Discourse ethics provided me with a theoretical lens for observing the relationship between explanations and perceptions of fairness. I adopted a discourse-ethics approach
because the associated theory focuses on the communication between stakeholders as the process by which ethical outcomes are achieved. Central to discourse ethics is the notion of an ideal speech situation, in which stakeholders engage as equals in the discourse, there is no coercion, and stakeholders have an opportunity to question the claims of others and to present their own claims and needs (Mingers & Walsham, 2010).

**The Importance of Discourse Ethics from the Explanation Perspective**

Discourse ethics can assist in showing why explanations can influence perceptions of fairness through components such as equal participation, balanced communication, and visible reasoning (Afrashteh, Someh, & Davern, 2020; Habermas, 1984). All of these components can nudge the discourse toward an ideal speech situation by affording transparency (Habermas, 1984; Mingers & Walsham, 2010).

**The Project’s Application of Discourse-Ethics Theory**

Equal participation requires a level playing field for interaction where the organization does not exploit a power differential to coerce the customer into providing data or making particular choices. However, because it is infeasible to engage every single customer in the decision-making process, making their perspective available to the organization’s decision-makers proves crucial. This enables decision-makers to look at things through customers’ eyes – seeing whether the decisions are truly fair to them rather than just in line with the organization’s point of view. When both of these perspectives are presented through explanations, the communication of how the decision was made is considered to be balanced. If balanced communication exists with regard to customers and organizations, a win–win situation can emerge for both, thereby leading to greater perceived fairness (Günther et al., 2017). In conjunction with bringing about balanced communication and leveling the playing field, the reasoning processes behind the algorithmic decision-making need to be rendered visible. Simultaneously, one must consider the extent to which the explanation method should expose the algorithm’s decision-making approach to the decision-makers and its manner of doing so, let’s confine them to the algorithm’s angle rather than create a space for them to apply their own judgment with regard to the two key perspectives, of customers and the organization.
Discourse ethics enables drawing attention to the influence of explanations on perceptions related to the fairness of decisions made by algorithms. Thus far, however, how explanations affect perceived fairness has gone largely unexplored. To examine how they do this, I used the other two tools: perspective-taking and decisional-guidance theory.

2.6.2. Perspective-taking

Decision-makers’ view of algorithmic decisions’ fairness stems from their sense of how discriminatory and biased the decision is relative to a specific individual. Since fairness could be perceived very differently by different stakeholders, with a decision that seems fair from a customer’s perspective perhaps not being perceived as fair by the company and vice versa, understanding the perceived fairness, which is heavily contingent on perspective, requires considering algorithmic decision-making from more than one stakeholder’s standpoint (Parker & Axtell, 2001). Therefore, drawing on perspective-taking theory, I framed explanations in a manner representing the organization’s and its customer’s standpoint both. This enabled me to explore how explanations affect the fairness decision-makers perceive when they consider the underlying logic from the respective stakeholder’s perspective. Per discourse-ethics theory, perspective-taking can also steer the setting toward an ideal speech situation, since the perspective of both main stakeholders in the process of decision-making is presented to the decision-makers, for input to exercising their own judgment accordingly.

2.6.3. Decisional Guidance vs. Restrictiveness

Decisional guidance is defined as “how a decision support system enlightens or sways its users” (Silver, 1991, p. 107). Among the many operationalizations for decisional guidance have been the use of multiple presentation formats (Wilson & Zigurs, 1999), manipulating explanation types (Limayem & DeSanctis, 2000), and provision of information about the accuracy of the system’s decisions (Montazemi, Wang, Nainar, & Bart, 1996). The literature distinguishes between two types of decisional guidance: informative and suggestive. The latter gives recommendations to the users as to which input to consider and how, whereas informative guidance provides relevant information to enhance decision-makers’ understanding without offering any suggested ways of
Chapter 2  The literature and conceptual development

consuming the information (Montazemi et al., 1996). The doctoral project focused on
the informative guidance to investigate its possible impact on the perception of fairness.
If one wishes to promote a sense of fairness, it is important to create space in which
decision-makers can exercise their judgment rather than guide them to a specific choice.

The organizations’ landscape has always manifested a tension between decisional
guidance and restrictiveness (Benbya et al., 2021). M. S. Silver (1990, p. 59) has
defined restrictiveness as “the degree to which and the manner in which a Decision
Support System limits its users’ decision making processes to a subset of all possible
processes.” Applying these definitions to the context of algorithmic decision-making, I
have been able to highlight the tension between explanations’ restrictiveness and
guidance in enabling or constraining decision-makers’ behavior and, more specifically,
in influencing their perceptions of decision by algorithms (Newport, 1990). Depending
on their content and the way they are framed and communicated with the decision-
makers, explanations can either restrict the decision-makers’ choices by limiting the set
of decision options or provide gentle guidance through creating a space for practicing
their own judgment after receiving an explanation. The former situation arises when the
explanations induce the decision-makers to focus merely on the piece(s) of information
provided while paying less attention to the information not presented in the explanations
(Asare & Wright, 2004; Silver, 1988). Less restrictive and more guidance-oriented
explanations, on the other hand, expose the algorithm’s decision strategies to a greater
degree and offer decision-makers more flexibility to interact with those strategies.

2.7. The Gaps in Understanding

The under-researched areas identified in the literature review and addressed by my
project are associated primarily with explanations for algorithmic decision-making, the
sense of the fairness represented by algorithms’ decision, and how explanations of
particular types influence a decision-maker’s perceptions of fairness.

While research into intelligent systems and explainable AI has explored the design of
explanation facilities, its focus has been primarily on improving the decision-making or
on human decision-makers’ reliance on the systems (Gregor, 2001; Tan et al., 2012;
Wang & Benbasat, 2007), while the impact of explanations on decision-makers’
perceptions of fairness in the context of algorithmic decision-making has yet to be explored in any depth. To examine that impact, one must identify what type of meaningful information is required for the explanations. What information has to be communicated to the organizations’ decision-makers depends on the reasons for using explanations (organizational objectives), the level of the decision-makers’ expertise, and the necessary scope of understanding of the decision. The literature has provided a starting point by exploring various kinds of explanation (Gregor & Benbasat, 1999; Wachter et al., 2017), but it falls short of examining how these affect decision-makers’ perceptions of fairness within the context of algorithmic decision-making (Benbya et al., 2021).

In that decision-making context, application of the concept of perspective-taking has consisted mainly of conceptualizing the ethics issues of employing algorithms and the consequences they might create for each stakeholder (Someh, Davern, Breidbach, & Shanks, 2019; Zuboff, 2015). Scholars have limited this attention to identifying the underlying ethics issues, without addressing them. However, drawing on discourse-ethics theory in the context of the relationship between explanations and perceived fairness highlights the importance of adopting multiple stakeholders’ perspective when communicating the explanations, so as to promote a speech situation that contributes to a sense of greater fairness. Accordingly, I adopted the perspective-taking-strategy to study how differences in the communicating of explanations framed through stakeholders’ perspective affect the decision-makers’ perception of fairness. This approach of taking multiple stakeholders’ perspectives provides more in-depth insight about the concerns of customers and organization as the key players in the context of this study (Clarke & Davison, 2020).

Notwithstanding the extensive research into the twin concepts of system restrictiveness and decisional guidance, along with their impact on the behavior of decision-support system, or DSS, users (Davern & Kamis, 2010; Goodwin, Fildes, Lawrence, & Stephens, 2011; Silver, 1988; Wheeler & Valacich, 1996), scholars have not explored the longstanding tension between these concepts in the context of algorithmic decision-making, let alone the related impact of explanations’ influence on perceptions.
of fairness While explanations can guide decision-makers to understanding the algorithm’s decision-making process, they can equally restrict exercising human judgment, by directing the focus toward the algorithm’s decision. Hence, I found fertile ground for examining the impact of restrictive explanation as a communication strategy on perceived fairness.

2.8. The Position of the Research Project

The most fundamental aim behind my research was to examine how explanations influence organizational decision-makers’ perceptions of fairness about the decision made by an algorithm. As discussed 2.4.1, the type of content supplied in the explanations can reflect different aspects of the algorithms, depending on the goal behind the provision of explanations and on the level of transparency required. With regard to the latter, the two most common goals for explaining algorithms’ decisions are to provide the users of the algorithms with an understanding of the decision and to generate trust by persuading of the decision’s validity (Miller, 2019). In response to these goals, firstly, I propose that understanding the system-generated decision requires information on how data support that decision. I refer to this as evidence of reasoning. This variable, which I examined in studies 2 and 3, can be structured in line with several logics, including factual and counterfactual reasoning. Secondly, to generate trust, increase the level of fairness perceived, and persuade of the algorithmic decision, one must explain and justify the performance of the algorithm. For systems’ users, metrics expressing evaluation of a given algorithm’s performance can shed light on the extent to which the decision made is correct and reliable. The variable connected with meaningful insight in this regard (“evaluation metrics”) was explored in studies 2 and 3. The next chapter addresses the various types of metrics that can be used and their possible differential impact. Together, the two variables cover both global and local types of explanations, per the taxonomy discussed in Section 2.4.1, evaluation metrics represent global-type explanation in that they articulate the performance of the system as a whole, while evidence of reasoning represents locally oriented explanation, addressing a single specific decision made on an individual customer at instance level.

The review of prior literature on explanations for algorithms’ decision-making and their relationship with perceptions of fairness provided theoretical grounding for the project
and supported answering the first sub-question, about identifying what constitutes the meaningful information that needs to be explained to organizations’ decision-makers. It was on this basis that, in line with the aim for the project, I identified evaluation metrics and evidence of reasoning as the information that must be communicated to the decision-makers if one wishes to enhance their sense of fairness. How this information should be framed and presented for making the greatest impact in this regard required empirical research, presented in the following chapters.
Chapter 3

ENHANCING THE PERCEIVED FAIRNESS OF ALGORITHMIC DECISION-MAKING VIA PERSPECTIVE-TAKING

This chapter presents the first experimental study, with which I focused on examining the impact of perspective-taking on organizational decision-makers’ perception of fairness in relation to explanation of algorithmic decision-making. The first section of the chapter introduces the research question addressed by the first experiment and the importance of it. Perspective-taking holds particular significance in relation to perceived fairness since individuals who differ in perspective might perceive the fairness of the same incident differently from each other. In the third section, the research model associated with the first empirical study is described and explained. After elaboration on the research model and the hypothesis-development work, the research methods are discussed, together with the details of the research design and of the experiment-development process. Then, Section 4.8 lays out the results of the data analysis, with the central finding being that perspective-taking influences perceptions of fairness. The final section of the chapter discusses the results and characterizes the study’s conclusions.

1 Elements of the chapter have been published in the following article:
3.1. Orientation to Perspective-taking

Understanding perceptions of the fairness surrounding algorithmic decisions requires one to consider the decisions and underlying logic from the angle of the stakeholders influenced by the algorithm’s decisions (emphasized by Parker & Axtell, 2001), or perspective-taking. Defined as the “adoption of another person’s viewpoint” (Parker & Axtell, 2001), this has been acknowledged for its contribution to helping companies develop a capacity for better understanding and responding to customers’ needs, predict the consequences of actions with regard to customers, and enhance their welfare (Dickey, Burnett, Chudoba, & Kazmer, 2007). The literature shows that perspective-taking creates an emotional connection between the person taking the perspective and that perspective’s target, which produces a change in the perspective-taker’s behavior (Batson et al., 1997; Lindsey, Yun, & Hill, 2007; Massi Lindsey, 2005). Additionally, as noted in Section 2.6.2, above, presenting both perspectives of the immediate stakeholders in the process of decision-making (the customer and the organization) facilitates balanced communication and equal participation, vital elements of ethical discourse. Having found little prior work exploring the influence of perspective-taking on organizational decision-makers’ perceptions of fairness, despite perspective-taking’s clear benefits, I sought to address the following research question:

• How does perspective-taking in the communication of algorithm’s decision influence the fairness perceived by the organization’s decision-makers?

To answer this question, I considered several ways of framing the explanations in communication such that they represent the perspectives of both organizations and customers. To this end, I designed a 2x2 between-subjects experiment in which the provision of information about how data support the decision (evidence of reasoning) and on aspects of the algorithm’s performance (evaluation metrics) are manipulated. I hypothesized that these manipulations would prompt different perspective-taking stances by the decision-maker, in line with my argument that different types of evidence of the reasoning employed can engender adoption of either the organization’s or the customer’s perspective by the decision-maker. Likewise, emphasizing different evaluation metrics would be expected to prompt the decision-maker to adopt an alternative stakeholder perspective. Accordingly, I set out to explore how the distinct
communication strategy of the explanations (i.e., how the model’s performance is explained and the evidence of its reasoning) from a customer vs. organizational perspective and prompting of the decision-makers to take either of them influence the fairness they perceive.

3.2. Theoretical Background

3.2.1. The Development of Perspective-taking

The cognitive process of perspective-taking is conceptualized as considering the attitudes and view of someone else within the given situation and trying to comprehend that situation (Lee et al., 2018; Te'eni et al., 2001). Importantly, taking someone else’s perspective can result in behavior change that favors the person targeted by the perspective-taking (Galinsky et al., 2008).

There has been much scholarly debate about perspective-taking and its benefits (Boland & Tenkasi, 1995; Dickey et al., 2007; Grant & Berry, 2011). Perspective-taking has demonstrated its value for designing a successful communication medium in knowledge-based organizations (Boland & Tenkasi, 1995), and several noteworthy studies have explored the role of perspective-taking in an organization’s relationships with its customers. For example, adopting a customer perspective can expedite service representatives’ resolution of customers’ problems (Dickey, Burnett, Chudoba, & Kazmer, 2007), and it supports customer satisfaction with call-center encounters (Axtell, Parker, Holman, & Totterdell, 2007; Hennig-Thurau, 2004). Experiments in an IS context provide evidence that software-product managers who have adopted a customer perspective are more likely to deescalate their commitment to launching a product that has proven to be defective (Lee et al., 2018).

Prior research attests that perspective-taking induces an emotional reaction (Batson et al., 1997; Lindsey et al., 2007; Massi Lindsey, 2005). By facilitating understanding of how the perspective’s target feels and thinks, it results in a change in behavior on the perspective-taker’s part. The change in behavior may arise in response to a sense of guilt foreseen by the perspective-taker. Individuals experience a sense of guilt under
three conditions: when they have become aware of a potential risk, found out about a potential solution for tackling the risk, and realized that someone else’s action could eliminate the risk. In adopting another’s perspective, an individual is more likely to experience this sense of guilt and has a tendency to act in such a way as to eliminate the source of that feeling (Batson et al., 1997). While researchers have studied related ethics consequences for various stakeholders in the algorithmic decision-making process (Someh et al., 2019; Zuboff, 2015), the potential of prompting perspective-taking so as to elicit behavior change in this context has not been empirically examined before.

The theory identifies three distinct types of perspective-taking: perceptual, cognitive, and affective. The first of these involves how people see things in their physical environment and occurs through the visual and auditory senses. If the intent is instead to understand how others feel, rather than sense, in a particular situation, affective perspective-taking would be needed. Finally, cognitive perspective-taking is related to intellectually understanding what others would think in that situation (Gasiorek & Ebesu Hubbard, 2017). Given my emphasis on getting decision-makers to think as if they were a customer or organizational stakeholder in the use of algorithmic decision-making, I focused on the last of these.

As for the nature of perspective-taking, it gets conceptualized variously as a process, ability, or trait (Gasiorek & Ebesu Hubbard, 2017). The process-oriented approach is based on the cognitive processes and mental actions through which a person acquires the target’s perspective. In the ability framework, in contrast, the aim is to judge the extent to which the perspective-taker is psychologically capable of immersion in the target’s situation. Finally, exploring perspective-taking as a trait-related concept deals with how much individual people’s characteristics affect spontaneous perspective-taking.

I approached perspective-taking as a process, since the study focused on providing participants with cognitive-situation-specific information about the perspective target (a customer or organizational stakeholder) in order to prompt them to take that target’s perspective (Gasiorek & Ebesu Hubbard, 2017). Perspective-taking theory informed my decisions on which aspects of algorithms to consider and experimentally investigate (with regard to evidence of reasoning and the evaluation metrics presented) and how to manipulate these aspects so as to influence the outcomes of interest (fairness perceptions).
3.3. The Study’s Research Model

I designed the study’s overall research model, depicted in the Figure 3.1, and my application of theory to explain how communicating different forms of explanation through different strategies affects the level of fairness perceived by decision-makers. The study used two forms of explanation, explaining the model’s performance and providing evidence of its reasoning. For the model-performance explanations, the output was explained via various evaluation metrics, while the evidence of reasoning took the form of information about the decision-making and how the model arrived at its decision. Communicating both forms of explanation from the organizational versus customer perspective provides an opportunity to explore explanations’ influence on the perception of fairness. The following subsections explain the constructs within the research model in more detail, culminating in presentation of the study’s two hypotheses.

Figure 3.1: The model developed for the first empirical study

3.3.1. Decision Outcomes

I identified three decision-related outcomes that might be influenced by the explanation of an algorithm’s decisions: the decision-maker’s assessment of how satisfying, reasonable, and fair the decision is. In line with Welsh’s aforementioned definition of fairness as the feeling people experience when receiving their fair share of the benefits available, this pertains mainly to the decision itself, with its effects. In contrast,
Chapter 3 Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

perceived satisfaction and reasonability are connected mostly with the decision-making process and how individuals feel about the process through which the decision has been made: does it seem satisfactory/reasonable? While the outcome variable of interest is perceived fairness, I assessed the others also, to discriminate between perceived fairness and other relevant perceptions of the decision/process and to illuminate how clearly the decision-makers distinguished among fair, satisfying, and reasonable decisions.

3.3.2. Model-Performance Explanations

As discussed in the previous chapter, explaining the performance of the algorithm can stimulate insight into its decision’s accuracy level. This form of global explanation can also highlight sources of potential error in the algorithm’s decision-making performance, thereby ultimately influencing perceptions about fairness. Therefore, it is important to explore model-performance explanations as a factor in perceptions of fairness.

It is my contention that communicating the model-performance explanations in line with different perspectives and prompting the decision-makers to take the corresponding perspective may influence their perceptions of fairness. To operationalize the organization’s and customer’s perspective, I used evaluation metrics of various sorts as the content for the model-performance explanations.

There are many types of evaluation metric, each geared for measuring the performance of the system for particular aims and applications (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013). For example, the most commonly used metric for evaluating an algorithm’s generalization ability is accuracy. For clarity of exposition, accuracy and other measurements are frequently described in the context of a classification task. In this approach, which suited the task in my experiment well, the total number of data points correctly predicted by a classifier out of all the data points is calculated. The resulting figure describes the quality of the decision made (McNee, Riedl, & Konstan, 2006). Accuracy is a very popular metric mainly because of its ease of computation, which also renders it more understandable for people with little knowledge of computer science. That said, the accuracy metric is of questionable utility when employed to provide information in rarely encountered cases such as the “edge cases” used in our experiment (Hossin & Sulaiman, 2015), wherein the situation falls near decision boundaries so might be placed in the wrong class by the algorithmic classifier.
Among the other commonly used metrics are precision and recall, figures that involve the count of true/false positives and true/false negatives. In a binary-classification setting such as a decision to grant/refuse credit, a correct approval decision is a true positive, and a true negative would be a correct denial decision, while approval when credit should have been denied is a false positive and deciding to deny it when approval would have been correct is a false negative. Precision is defined as the number of true positives divided by the total number of positive identifications (true plus false positives). Recall, in contrast, is the number of true positives divided by the number of elements that should have been identified (i.e., true-positive plus false-negative results).

The choice of evaluation metrics influences how the performance of an algorithm is perceived. For example, consider COMPAS, the US parole DSS mentioned in the introductory chapter. Although the accuracy, precision, and recall of COMPAS were considered satisfactory, it was inappropriately discriminatory in the predictions of parolee recidivism it made for black defendants. On closer inspection, the false-negative rate for these defendants was excessive and biased. Clearly, the choice of evaluation metrics to focus on has consequences.

In considering the evaluation metrics available for explaining the performance of a model, tradeoffs swiftly become evident. The question, then, is what constitutes the appropriate tradeoff. I posit that the appropriate tradeoff is a function of the perspective taken by the decision-maker. By implication, emphasizing different metrics is consistent with a user adopting a different perspective. This matter is explored further in conjunction with the development of the research model, below.

Organizations are concerned primarily with the accuracy of the systems they employ, in the sense that highly accurate performance justifies the given system’s use for decision-making (McNee et al., 2006). Providing figures for the overall accuracy of its performance is, hence, likely to represent the organization’s perspective. However, these figures do not provide a full picture of the algorithm’s performance where customers are concerned. Considering other aspects of performance requires metrics additional to accuracy, such as the percentage of false negatives, false positives, true negatives, and true positives. These metrics afford more detailed analysis than does
merely the proportion of correct classifications (accuracy) and can be customized for the explanations that customers require.

Providing an error matrix that shows all four figures (false negatives, false positives, true negatives, and true positives) characterizes the decision landscape in a manner more consistent with a customer perspective (e.g., the false-negative rate portrays the extent to which the decision might deprive customers of a right to a higher credit limit). Importantly, the manipulation does not entail changing the algorithm’s accuracy or behavior; rather, it consists of adjusting which evaluation metric for a given algorithm is made prominent, in line with the following hypothesis:

**H1**: Decision-makers provided with an error matrix explaining the model’s performance are more likely to

a. Perceive the algorithm’s decision to be less fair than decision-makers given a metric for only accuracy in the model’s explanation

It is worth stating explicitly that the study did not assess possible change in ultimate decision-maker behavior (e.g., overruling the algorithm). The focus was on exploring perceptions about the choices presented.

### 3.3.3. Evidence of the Reasoning

Fundamentally, all types of explanations discussed above involve providing evidence of the reasoning employed: information about how data and analytics support the decision made by the algorithm (Miller, 2019). Proceeding from my synthesis of prior literature, I sought to examine local explanations and contrast the role of counterfactual versus factual evidence of reasoning, consistent with a customer versus organization perspective. The focus on local explanations here arose from their nature as inherently closer to an individual customer’s standpoint, as opposed to global explanations (which are, by nature, broader and thus more reflective of an organizational perspective). I chose counterfactual vs. factual reasoning because counterfactual explanations force a contrastive perspective on the recipient. Prior research in cognitive science suggests that when individuals are interested in understanding why a decision was not in their favor,
they seek contrastive explanations (Miller, 2019). The following example of evidence behind the reasoning for denying a credit-limit increase request is a case in point.

Factual reasoning: The customer’s request is denied because his monthly expenses total $2,300 and his net monthly pay is $3,100.

Counterfactual reasoning: The customer would have been approved if his monthly expenses had been less than $2,000 and net monthly pay were over $3,200.

For this study, I posited that framing the evidence of reasoning as local explanations anchored in different perspectives through its communication and prompting the decision-maker to take either of those perspectives may influence the sense of fairness. In this setting, I used factual vs. counterfactual reasoning to operationalize the organizational and customer perspective.

The counterfactual explanation provides a contrastive explanation that makes it clear from a customer perspective what would have been required for, in this case, approval of a higher credit limit. Since it is potentially more meaningful to customers, this counterfactual explanation is more likely to guide decision-makers toward a customer perspective. Conversely, factual explanations, providing information about the most important attributes on which the decision was based, are more likely to represent the organization’s perspective. Consequently, I hypothesized that employing counterfactual as opposed to factual explanations would influence the perception of fairness. This should be particularly salient with regard to edge cases, in which the decision is somewhat ambiguous and the system was close to recommending an alternative decision. Specifically, I hypothesized thus:

**H2**: Decision-makers provided with counterfactual explanations are more likely to

a. Perceive algorithm-generated decision to be less fair

than are decision-makers given factual explanations
Chapter 3  Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

3.4. The Research Method

3.4.1. Participants

The experiment gave participants information about a credit application and asked them to determine, in light of the decision by the algorithm, whether or not the customer’s request for a higher credit limit should be approved. All participants were native English-speakers over the age of 25 who were registered with the Prolific.co online subject-pool platform. For greater external validity, the project used preliminary filtering to include only people who had earned at least an undergraduate degree and had management experience in the financial industry. To confer sufficient power for testing the hypotheses, I recruited 200 people to take part (with 50 per cell in a 2×2 design).

The data from five participants were discarded because of either unreasonably rapid completion of the task or demonstration of insufficient recall/understanding of the experiment conditions. The criterion for the latter was conclusive failure of two manipulation checks (using two questions). Of the 195 participants remaining, 105 were male and 90 were female. As for education, 121 had an undergraduate degree, 72 a sub-doctoral graduate qualification, and two a doctoral degree (a PhD). Most were in their mid-thirties. On average, the subjects were proficient in data analysis, with an average of 2.7 years’ experience in making credit decisions. Table 3.1 presents the participants’ background data.
Table 3.1: Participant characteristics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>35.6 years</td>
<td>19.19</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>90</td>
</tr>
<tr>
<td>Level of education</td>
<td>Undergraduate</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>Graduate studies</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Doctoral degree</td>
<td>2</td>
</tr>
<tr>
<td>Data-analytics experience</td>
<td>4.01</td>
<td>1.863</td>
</tr>
<tr>
<td>Experience in making credit decisions</td>
<td>2.70</td>
<td>2.319</td>
</tr>
</tbody>
</table>

Sample size = 195. No missing data.

Both data-analytics experience and credit-decision experience are on a seven-point scale with the anchors 1 = none and 7 = substantial.

3.4.2. The Task and Procedures in the Experiment

A 2x2 between-subjects experiment design was employed, contrasting factual vs. counterfactual reasoning and presentation of the system’s performance accuracy vs. that of an error matrix for its performance. The dependent variables were perceptions of the satisfaction, reasonability, and fairness of the system’s performance and of its decision.

A pilot test with 16 people was conducted before the actual experiment, to make sure that the process was understandable and the case scenario was appropriate. The pilot phase provided participant feedback that was valuable for the experiment and gave a sense of the time needed for the decision-making task: about 10 minutes.

Several variables for inter-individual differences with potential to confound the results were included as control variables (these are discussed in connection with the experiment material, in the next subsection). The setting was a Web-based one utilizing the Qualtrics survey tool and applying a scenario-based method, as commonly employed in ethics and social-psychology research. Research has shown that the way
people act in scenario-based experiments is consistent with how they act throughout their life in the real world (Woods, Walters, Koay, & Dautenhahn, 2006).

The experiment involved a simulated credit-limit increase decision by a manager at a financial institution using an analytics system that makes a recommendation as to whether the increase request should be approved. The purpose behind this design and its manipulation of the evaluation metrics and evidence of reasoning was to prompt the participants in various ways to take the customer’s perspective and, thereby, examine whether taking another stakeholder’s perspective produces greater awareness of some possible unintended consequences of algorithmic decision-making, and whether any such change ultimately affects the decision-maker’s perceptions of the decision. The scenario was designed to represent an edge case – a customer profile lying near the decision boundaries and therefore with potential to be sorted into the wrong class by the algorithmic classifier. A denial decision is made by algorithms for the specific case of this scenario based on the customer’s profile.

**Operationalization of the Independent Variables**

The operationalization of the independent variables drew from perspective-taking theory (Gasiorek & Ebesu Hubbard, 2017). With the setting presented above, I expected an emphasis on different evaluation metrics to yield different outcomes (Afrashteh, Davern, & Asadi Someh, 2020), since presenting the accuracy of the system’s performance represents the organization’s perspective while the customer’s is framed through provision of the error matrix. Again, the matrix compares the algorithm’s classification decisions with objective targets, with the presence of information on the false-negative rate (the percentage of the customers whose application for a credit increase was incorrectly denied) being likely to lead to perceptions more consistent with a customer perspective than an organizational one. Taking a customer perspective helps a human decision-maker assess whether the analytics system’s performance and the algorithm’s decision yield unfair consequences for the customer. With such a facility, when identifying such a case, they could override the algorithmic decision with a decision they perceive as fairer.
3.4.3. Materials for the Experiment

The participants were provided with four scenarios, with the type of explanation for the reasoning and the communication of the algorithm’s performance evaluation both manipulated with regard to perspective. Table 3.2 shows the number of participants exposed to each set of conditions.

<table>
<thead>
<tr>
<th>Evidence of reasoning</th>
<th>Factual</th>
<th>Counterfactual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>47</td>
<td>40</td>
</tr>
<tr>
<td>Error matrix</td>
<td>58</td>
<td>50</td>
</tr>
</tbody>
</table>

The dependent variables assessed in this study were participant perceptions of satisfaction, reasonability, and fairness. Each of these was measured in two steps: asking about perceptions of the system’s performance (the process), then about perceptions as to the algorithm’s decision. All dependent variables were measured on a six-point semantic differential scale, since it was critical not to have participants biased in either direction (Osgood, 1964). Also, I used direct measurement to address the phenomenon at large: whether or not the participant chose to override the denial decision offered by the analytics. I employed binary coding for the latter choice: the decision was coded as “1” if the participant overrode the decision, and “0” was assigned if the participant conformed with the analytics decision.

**Control Variables**

People’s trust in algorithms and their performance is among the most important factors influencing their views of the fairness of decisions by algorithms (McKnight, Carter, & Clay, 2009). McKnight et al. (2009) identified two dimensions of trust in technology:
Chapter 3 Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

“faith in general technology” and “trusting stance – general technology.” I found the former, faith in technology in general, suitable for controlling the influence that trust in algorithms may exert on fairness perceptions in that it refers to “individuals’ beliefs about attributes of Information Technology in general” rather than focusing on the degree to which they believe in a reliable positive outcome of technology (McKnight et al., 2009, p. 8). Since people may differ in the level of faith they have in algorithms, an impact on their perceptions of the system and the decision made by analytics may well follow (McKnight et al., 2009). It makes a clear difference if some people put blind faith in algorithms and refuse to question their performance.

Another control variable addressed open-mindedness. Referred to here as actively open-minded thinking (AOM), the extent to which individuals are open to questioning initially preferred outcomes varies. This could have an impact on participants’ cognitive engagement with the experiment’s task. Moreover, using this variable as a control may address participants’ “myside bias,” defined as “the tendency to thinking ways that strengthen whatever possible conclusions are already strong” (Baron, Scott, Fincher, & Metz, 2015, p. 267).

The discriminant validity (distinctness) of the control variables was assessed through exploratory factor analysis. The key results, reported in Table 3.3, can be summarized

<table>
<thead>
<tr>
<th>Variable's name</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOM 1</td>
<td>0.20</td>
<td>0.272</td>
<td>-0.153</td>
<td>0.760</td>
</tr>
<tr>
<td>AOM 2</td>
<td>0.20</td>
<td>0.652</td>
<td>-0.200</td>
<td>0.212</td>
</tr>
<tr>
<td>AOM 3</td>
<td>0.068</td>
<td>-0.041</td>
<td>0.773</td>
<td>0.165</td>
</tr>
<tr>
<td>AOM 4</td>
<td>0.086</td>
<td>0.605</td>
<td>0.080</td>
<td>0.085</td>
</tr>
<tr>
<td>AOM 5</td>
<td>0.147</td>
<td>0.336</td>
<td>-0.498</td>
<td>0.176</td>
</tr>
<tr>
<td>AOM 6</td>
<td>-0.082</td>
<td>0.618</td>
<td>0.427</td>
<td>0.079</td>
</tr>
<tr>
<td>AOM 7</td>
<td>0.032</td>
<td>0.574</td>
<td>-0.143</td>
<td>-0.588</td>
</tr>
<tr>
<td>AOM 8</td>
<td>-0.105</td>
<td>0.539</td>
<td>-0.262</td>
<td>-0.122</td>
</tr>
<tr>
<td>AOM 9</td>
<td>0.062</td>
<td>0.089</td>
<td>0.746</td>
<td>-0.180</td>
</tr>
<tr>
<td>AOM 10</td>
<td>0.102</td>
<td>0.721</td>
<td>-0.007</td>
<td>-0.018</td>
</tr>
<tr>
<td>FGT 1</td>
<td>0.857</td>
<td>-0.040</td>
<td>-0.014</td>
<td>-0.019</td>
</tr>
<tr>
<td>FGT 2</td>
<td>0.870</td>
<td>-0.044</td>
<td>0.018</td>
<td>-0.038</td>
</tr>
<tr>
<td>FGT 3</td>
<td>0.843</td>
<td>0.046</td>
<td>-0.035</td>
<td>-0.002</td>
</tr>
<tr>
<td>FGT 4</td>
<td>0.845</td>
<td>0.035</td>
<td>0.076</td>
<td>0.021</td>
</tr>
</tbody>
</table>

thus: four components were extracted, together explaining 60.36% of the variance. Of the factors covered by this cumulative percentage, the rotated component matrix (in Table 3.4, below) shows that faith in technology in general (FGT) has the strongest factor loading, displaying values above 0.8, with all four of its items loaded together. I performed confirmatory factor analysis to assess the measurement properties of the instruments addressing the control variables. Firstly, the instruments’ internal consistency was assessed. For actively open-minded thinking, the Cronbach’s alpha was 0.12, which did not meet the standard criteria; therefore, this variable was excluded from our controls. For faith in general technology, the corresponding value was 0.89. This was above the threshold, so the construct was deemed internally consistent (Barclay, Higgins, & Thompson, 1995; Fornell & Larcker, 1981). Then, the loading of each instrument item on its corresponding factor was assessed for the faith-in-technology variable, for checking the reliability of the items. The loadings of the items met the criteria: they were all above 0.7, which is considered excellent and indicates a satisfactory result for the factor analysis (Tabachnick & Fidell, 2001).

The instruments included asking the participants to justify their responses. I subjected the justification details to thorough data analysis to examine what constituted the main reason behind the final decision they made when overruling that offered by the system.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Measure</th>
<th>Std.loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGT 1</td>
<td>I believe that most algorithmic decision making systems are effective at what they are designed to do.</td>
<td>0.791</td>
</tr>
<tr>
<td>FGT 2</td>
<td>A large majority of algorithmic decision making systems are excellent.</td>
<td>0.810</td>
</tr>
<tr>
<td>FGT 3</td>
<td>Most algorithmic decision making systems have the features needed for their domain.</td>
<td>0.796</td>
</tr>
<tr>
<td>FGT 4</td>
<td>I think most algorithmic decision making systems enable me to do what I need to do.</td>
<td>0.809</td>
</tr>
</tbody>
</table>

*The SPSS Statistics 27 was used for the factor analysis. Extraction method: maximum-likelihood, with no rotation since only one factor was explored.*
Chapter 3 Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

The results show that contextual information describing the health circumstances of the customer in question had a strong influence on the participants. It was associated with adopting the customer’s perspective and also showed a connection with satisfaction, reasonability, and fairness perceptions related to the decision offered by analytics. This impact can be explained by participants deeming health-related issues highly important. On this basis, a new variable was defined for the analysis, called “context-awareness.” This was used as a covariate for statistically addressing the impact the context-specific information on the customer had on the dependent variables. In this case, the instances of considering the contextual information on the reason for the customer’s late payment were assigned the code “1” to refer to the participant’s awareness of said context. For participants who did not take the customer-related contextual information into account, “0” was assigned, to reflect not being swayed by it.

3.5. Data Analysis and Results

3.5.1. Descriptive Data

The descriptive statistics for the dependent variables related to fairness and for the participants’ decisions relative to those recommended are provided in Table 3.5 and Table 3.1, respectively. The tables present the data after pre-analysis screening.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Mean</th>
<th>Median</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reasonability (of the system’s performance)</td>
<td>4.47</td>
<td>5.00</td>
<td>1.14</td>
</tr>
<tr>
<td>Satisfaction (with the system’s performance)</td>
<td>4.46</td>
<td>5.00</td>
<td>1.15</td>
</tr>
<tr>
<td>Fairness (of the system’s performance)</td>
<td>4.25</td>
<td>4.00</td>
<td>1.22</td>
</tr>
<tr>
<td>Reasonability (of the algorithm’s decision)</td>
<td>4.16</td>
<td>4.00</td>
<td>1.28</td>
</tr>
<tr>
<td>Satisfaction (with the algorithm’s decision)</td>
<td>4.07</td>
<td>4.00</td>
<td>1.40</td>
</tr>
<tr>
<td>Fairness (of the algorithm’s decision)</td>
<td>3.88</td>
<td>4.00</td>
<td>1.40</td>
</tr>
</tbody>
</table>

All measurements are on a six-point semantic differential scale with anchors 1 = negative pole and 6 = positive pole.
In all, 200 participants completed the entire experiment, with all response sets being subjected to a manipulation check for gauging the extent of the participant’s awareness of the presence of error matrix and the explanation logic of the material presented to them. I conducted 2×2 factorial analysis of covariance (ANCOVA) to examine the effects of the forms of evaluation metrics and evidence of reasoning on perceptions of satisfaction with the system’s performance and with the decision made by the algorithm after accounting for the effects of the faith-in-technology covariate. Though the final decision by the organizational decision-maker is not the outcome of central interest, the frequencies of contrary decisions in each condition are presented in Table 3.6. There is a clear pattern: more of the participants (nearly 76%) overrode the denial decision to approve the request when provided with an error matrix and factual reasoning in combination, whereas 66% of the participants given a matrix accompanied by counterfactual reasoning approved the non-recommended credit-limit increase.

3.5.2. Manipulation Checks

On average across the treatments, 69.50% of the participants demonstrated accurate recall of the evaluation-metrics manipulation, and 92.4% recalled the manipulation of
the evidence for the reasoning correctly. Overall, the average rates are high enough for one to state that the independent variables were well-designed and acceptably recalled.

3.5.3. Correlations

Some interesting patterns are visible from the correlation matrix presented in Table 3.7. The presence of evaluation metrics had a significant negative impact on the satisfaction, reasonability, and fairness perceived in connection with the system’s performance, at $p = 0.01$ level. The same impact is visible for perceived satisfaction with the decision made by the system ($p = 0.05$). Moreover, the evidence of reasoning significantly influenced perceived satisfaction with the decision, at $p = 0.05$ level. Prior experience with data analytics correlated positively with the sense of decision satisfaction, along with the positive correlation with ultimate decision and context-awareness. Likewise, experience specifically with making credit decisions showed a positive correlation with perceived decision reasonability, satisfaction, and the final decision, at 0.05 level. The more experienced the decision-makers were, the more likely they were to consider the customer’s contextual information. Prior experience with analytics, context-awareness, and faith in technology were considered as possible covariates, with the correlation patterns showing that the last of these significantly affected all dependent variables except perceived decision fairness. Therefore, it was considered a covariate throughout the further analysis. On account of its meaningful significant negative correlation with perceptions of satisfaction with the decision, context-awareness was treated as a covariate in analysis of some dependent variables. Finally, and perhaps not surprisingly, greater satisfaction with the recommended decision was negatively correlated with overriding a decision to switch it from denial to approval. Detailed information on the patterns can be found in Tables 3.7, 3.8, and 3.9.
### Table 3.7: Correlation matrix for the variables

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Evaluation metrics</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Evidence of reasoning</td>
<td>0.001</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Perceived reasonability (of system performance)</td>
<td>-0.275** -0.008</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Perceived satisfaction (with system performance)</td>
<td>-0.191** -0.036 0.771**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Perceived fairness (of system performance)</td>
<td>-0.370** -0.041 0.760 0.674**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Perceived reasonability (of decision by analytics)</td>
<td>-0.153 0.109 0.149 0.771** 0.165**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Perceived satisfaction (with decision by analytics)</td>
<td>-0.141** 0.154** 0.070 0.060 0.177** 0.819**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Perceived fairness (of decision by analytics)</td>
<td>-0.116 0.089 0.071 0.019 0.170 0.794** 0.812**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Final decision</td>
<td>0.114 -0.091 0.020 0.042 -0.084 -0.488** -0.602** -0.518**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Faith in technology</td>
<td>-0.117 0.076 0.215** 0.230** 0.253** 0.231** 0.242** -0.020 0.173**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Awareness of context</td>
<td>0.075 -0.055 0.037 0.016 0.008 -0.285** -0.347** -0.308** 0.035 -0.296**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Data-analysis experience</td>
<td>0.008 0.086 0.009 -0.006 -0.001 0.129 0.152** 0.008 0.148** -0.082 0.144**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Experience with credit decisions</td>
<td>0.037 0.082 0.108 0.110 0.078 0.173** 0.163** -0.056 0.151** -0.065 0.449** 0.219**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Levels of p-values: * = \( p \leq 0.05 \), ** = \( p \leq 0.01 \), and *** = \( p \leq 0.001 \).*
3.5.4. Analysis of Covariance

ANCOVA was applied to examine the effects of the model-performance explanation and decision-making reasoning framed through the lens of stakeholder perspective on perceived satisfaction, reasonability, and fairness related to the system’s performance.

Table 3.8 shows that the perceptions of the reasonability of system performance varied between when evaluation metrics were provided and when they were not. This effect is significant ($p = 0.000$). The aforementioned analysis explains nearly 10% of this difference in the dependent variable; the adjusted $R^2$ value is 0.099. The presence of evaluation metrics showed a significant negative effect on the perceived reasonability of the system’s performance: presenting the evaluation metrics decreased the reasonability perceived by the participants with regard to the system’s performance.

Table 3.9 shows that satisfaction with the system’s performance varied between the condition in which the evaluation metrics were provided and that in which they were not. The effect is significant, with $p = 0.018$. Around 7% of the associated difference in the dependent variable is explained by this analysis; adjusted $R^2 = 0.071$. The presence of evaluation metrics had a significant negative effect on the satisfaction expressed with the system’s performance also. That is, their presence produced less participant-perceived satisfaction with its performance.

Table 3.10 suggests that the perceived fairness of system performance likewise varied between when the evaluation metrics were and were not provided. This effect too is significant ($p = 0.000$): 17% of the difference detected in the relevant dependent variable is explained by this analysis (adjusted $R^2 = 0.170$). The presence of evaluation metrics had a significant negative impact on the sense of the fairness of the system’s performance; i.e., their provision was associated with lower perceived fairness levels.

No significant influence of error-matrix provision or counterfactual explanations on the perceived reasonability of the decisions made by algorithms emerged in Table 3.11.

Table 3.12, in turn, shows that satisfaction with the decision made by the algorithm differed between subjects provided with factual reasoning and those supplied with counterfactual reasoning. This effect is significant, with $p = 0.05$. The analysis accounts for around 17% of the difference in the dependent variable (adjusted $R^2 = 0.176$). The use of a factual explanation had a significant negative effect on perceived satisfaction with
the analytics-produced decision; i.e., provision of factual reasoning created less satisfaction. In the analysis of satisfaction perceptions related to the decisions, context-awareness was treated as a covariate with faith in technology. The reason is that the initial correlation statistics suggested that a significant negative correlation exists between these two variables. Moreover, when it was examined as a possible covariate, with its effect hence normalized over all participants, a significant impact of the evidence of reasoning on perceived decision satisfaction emerged.

Table 3.13 shows that no significant influence of error matrix or counterfactual explanations’ presentation on the perceived fairness of the algorithmic decisions was evident. Finally, although the results in this regard were not statistically significant, factual explanations displayed a negative effect on the perceived fairness of a given decision made by analytics. This pattern’s emergence is discussed in more depth in Subsection 3.6.

Table 3.8: ANCOVA results where the dependent variable is perceived reasonability with the system’s performance

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>16.516</td>
<td>16.516</td>
<td>13.952</td>
<td>0.000</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>8.386</td>
<td>8.386</td>
<td>7.084</td>
<td>0.008</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>0.861</td>
<td>0.861</td>
<td>0.705</td>
<td>0.402</td>
</tr>
<tr>
<td>Error</td>
<td>191</td>
<td>226.106</td>
<td>1.184</td>
<td>0.538</td>
<td>0.464</td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>4,154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Adjusted R² = 0.099.*

Table 3.9: ANCOVA results where the dependent variable is perceived satisfaction with the system’s performance

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>7.185</td>
<td>7.185</td>
<td>5.730</td>
<td>0.018</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>11.265</td>
<td>11.265</td>
<td>8.985</td>
<td>0.003</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>0.118</td>
<td>0.118</td>
<td>0.094</td>
<td>0.759</td>
</tr>
<tr>
<td>Error</td>
<td>191</td>
<td>239.482</td>
<td>1.254</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>4,142.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Adjusted R² = 0.071.*
Chapter 3  Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

Table 3.10: ANCOVA results where the dependent variable is perceived fairness with the system’s performance

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>34.431</td>
<td>34.431</td>
<td>27.678</td>
<td>0.000</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>12.833</td>
<td>12.833</td>
<td>10.316</td>
<td>0.002</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>0.200</td>
<td>0.200</td>
<td>0.161</td>
<td>0.689</td>
</tr>
<tr>
<td>Error</td>
<td>191</td>
<td>237.597</td>
<td>1.244</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>3,815.00</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.170$.

Table 3.11: ANCOVA results where the dependent variable is the level of perceived reasonability with the algorithm’s decision

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>3.756</td>
<td>3.756</td>
<td>2.623</td>
<td>0.107</td>
</tr>
<tr>
<td>Evidence of reasoning</td>
<td>1</td>
<td>2.056</td>
<td>2.056</td>
<td>1.436</td>
<td>0.232</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>12.217</td>
<td>12.217</td>
<td>10.101</td>
<td>0.004</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>22.522</td>
<td>22.522</td>
<td>15.729</td>
<td>0.000</td>
</tr>
<tr>
<td>Error matrix × evidence of reasoning</td>
<td>1</td>
<td>0.474</td>
<td>0.474</td>
<td>0.331</td>
<td>0.566</td>
</tr>
<tr>
<td>Error</td>
<td>189</td>
<td>270.623</td>
<td>1.432</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>3,691.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.127$.

Table 3.12: ANCOVA results where the dependent variable is the level of perceived satisfaction with the algorithm’s decision

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>4.992</td>
<td>4.992</td>
<td>2.743</td>
<td>0.099</td>
</tr>
<tr>
<td>Evidence of reasoning</td>
<td>1</td>
<td>7.061</td>
<td>7.061</td>
<td>3.880</td>
<td>0.050</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>17.624</td>
<td>17.624</td>
<td>9.684</td>
<td>0.002</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>40.668</td>
<td>40.668</td>
<td>25.192</td>
<td>0.000</td>
</tr>
<tr>
<td>Error matrix × Evidence of reasoning</td>
<td>1</td>
<td>0.010</td>
<td>0.010</td>
<td>0.006</td>
<td>0.941</td>
</tr>
<tr>
<td>Error</td>
<td>190</td>
<td>345.775</td>
<td>1.820</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>3,605.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.176$. 

Table 3.10: ANCOVA results where the dependent variable is perceived fairness with the system’s performance

Table 3.11: ANCOVA results where the dependent variable is the level of perceived reasonability with the algorithm’s decision

Table 3.12: ANCOVA results where the dependent variable is the level of perceived satisfaction with the algorithm’s decision
Table 3.13: ANCOVA Results (Dependent Variable: Perception of Fairness about Algorithm’s Decision.

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error matrix</td>
<td>1</td>
<td>3.475</td>
<td>3.475</td>
<td>1.909</td>
<td>0.169</td>
</tr>
<tr>
<td>Evidence of the reasoning</td>
<td>1</td>
<td>2.103</td>
<td>2.103</td>
<td>1.156</td>
<td>0.284</td>
</tr>
<tr>
<td>FGT (covariate)</td>
<td>1</td>
<td>13.416</td>
<td>13.416</td>
<td>9.684</td>
<td>0.002</td>
</tr>
<tr>
<td>Context-awareness</td>
<td>1</td>
<td>30.929</td>
<td>30.929</td>
<td>16.992</td>
<td>0.000</td>
</tr>
<tr>
<td>Error matrix × evidence of the reasoning</td>
<td>1</td>
<td>0.024</td>
<td>0.024</td>
<td>0.013</td>
<td>0.909</td>
</tr>
<tr>
<td>Error</td>
<td>190</td>
<td>345.775</td>
<td>1.820</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>385.046</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2$: 0.102

3.6. Discussion and Conclusions

Drawing on what the intelligent-systems and perspective-taking literature has revealed about explanations, this study theoretically articulated and empirically tested the impact of model-performance explanations and evidence of reasoning framed from the organization’s and from the customer’s perspective on organizational decision-maker’s perceptions of satisfaction, reasonability, and fairness related to the system’s performance and the decision put forth.

The core reason behind choosing evaluation metrics and evidence of reasoning is that these explain the two important aspects of the algorithms: evaluation metrics explain the performance of the system, and evidence of reasoning explains the system’s logic, how it reached the decision. Additionally, the two cover both aspects of scope – evaluation metrics are regarded as global explanations and evidence of reasoning as a local one. Studying the impact of both local and global explanations on perceptions of fairness yields more comprehensive understanding, better addressing the research question.

The aim with these manipulations was twofold: 1) to get the decision-makers to apply their own judgment about the decision made for an edge-case scenario by prompting them to take the perspective of both the organization and customers and 2) to explore whether doing so affects the perceived fairness of the decision made by the algorithm.
Chapter 3  Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

The experiment followed a two-phase design logic, so the discussion below is organized accordingly.

The first stage of the experiment was intended to emphasize the important role of the perspective-taking and, thereby, reveal its role in perceptions of the system’s performance. Through the evaluation metrics presented, the set of conditions represented the organization’s and a customer’s perspective. Again, attention to the overall accuracy of the system’s performance captured the organization’s perspective (in that organizations are concerned primarily about the accuracy of the systems they employ, with higher overall accuracy of the system’s performance justifying the use of algorithmics for decision-making). Taking the customer’s perspective in communicating model performance explanations, a full error matrix needs to be provided rather than just an overall-accuracy figure, thus highlighting the false-negative rate.

The second stage exposed the decision-maker to a specific case of a customer, which explicitly triggered taking the customer’s perspective by means of contextual information about the customer’s health circumstances, related to a car accident that had resulted in a late payment. The conditions presented after this framed two types of evidence of reasoning, again representing both an organizational and a customer perspective. Here, factual reasoning represented the former, pointing out the most important attributes on which the system’s decision was based. In contrast, the condition with counterfactuals accentuated the customer’s perspective, by illustrating the smallest possible changes required for a customer to obtain the desired result (credit approval in this case). The aim for this stage was to examine whether there are differences between reasoning types in the ensuing perceptions about a given decision made by analytics.

The first stage of the experiment was designed in accordance with the hypothesis that providing evaluation metrics would influence the participants’ perceptions about the performance of the system. The results indeed attest that the presence of an error matrix decreased the perceived satisfaction, reasonability, and fairness related to system performance and reduced the perceived reasonability of the decision generated by analytics. This result can be explained by the error matrix’s role in calling the decision-maker’s attention to certain limitations of the model, thus highlighting the likelihood of incorrect decisions to deny credit to customers whose requests should have been approved (represented by the false-negative rate).
Chapter 3    Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

The second stage addressed the hypothesis that decision-makers who are supplied with counterfactual explanations are more likely to perceive a decision made by analytics as less fair and satisfying to customers. Contrary to the hypothesis, however, presenting factual reasoning was associated with less perceived satisfaction with the algorithm’s decision. Furthermore, providing counterfactual reasoning functioned counter to the negative-perception-inducing effect of presenting an error matrix: participants were less satisfied with the decision when they were provided with an error matrix and factual explanation. A similar pattern emerged for perceptions of its fairness, but this was not statistically significant. The pattern’s weakness might stem from the fact that “fairness” is a loaded and multifaceted term. Participants may have found it difficult to judge fairness with the level of information and evidence they were given. That said, since perceived fairness specific to the system’s performance did show a significant pattern, one can conclude that the concept of fairness might be deemed less problematic when not connected with a specific case.

Counterfactual reasoning can help a decision-maker identify the criteria as not being met, by clarifying the decision boundaries such that the decision may be satisfying. Therefore, such reasoning could function to justify the decision to decision-makers, thereby leaving a more positive impression of it. Per the literature, counterfactual explanations are among the best insight-capturing types of reasoning with regard to how algorithms arrive at a decision (Miller, 2019; Wachter et al., 2017). The results of my study demonstrate that counterfactual reasoning can serve as useful justification in cases involving a well-performing analytics system that people believe yields the best outcome. This role in supporting acceptance can cut both ways, however. When the system’s performance is plagued by potential for error, as reflected in a high rate of false negatives and false positives, the counterfactuals’ explanation of the decision might exacerbate the issue. They may constrain the situation by showing the cutoffs and closing off room for interpretation of the system’s decision. This factor may lie behind the finding that counterfactuals are not helpful for internal review processes (since they do not aid in forming a skeptic’s stance for reviewing the system with edge cases in mind). Notwithstanding this issue, counterfactual reasoning is preferable for communication with customers, because it not only describes the decision parameters.
Chapter 3  Enhancing the perceived fairness of algorithmic decision-making via perspective-taking

but also specifies a target pointing to what they should do if they want to receive a different decision.

The results of the first experiment demonstrate that the concept of room for interpretation might have an impact on perceived satisfaction and fairness. If organizations’ decision-makers can be steered toward assessing whether there is room for interpreting the system’s decision rather than simply letting it make all decisions itself, they will grow more skeptical. In a real-world setting, no decision made by an algorithm is rock-solid: no system is 100% accurate. The results produced by analytics are always “fuzzy,” even their explanations. Yet the first experiment showed that the explanations were taken as-is, as if they were perfect because they could not cognitively engage the decision-makers and create the space for their judgement about the algorithm’s decision-making and its decision outcome. To articulate the room for interpretation and encourage a skeptical approach by the decision-makers, the setting needs to make them aware that the explanations are not perfect and are noisy. Moreover, they need to recognize that the domain of credit decisions too is fuzzy, in various sorts of circumstances.

The matter of the room for interpretation brings the tension between decisional guidance and restrictive explanations to the fore (Benbya et al., 2021; Silver, 1991). Exploring the extent to which counterfactual explanations, in order to yield the greatest benefit, should be guiding rather than restrictive should advance understanding of how room for interpretation can be created, room that influences perceptions of fairness in consequence.

With this aim, I carried out a second experiment, in which I manipulated the explanation’s restrictiveness by means of interactivity provision. The second empirical study, designed to examine whether perceived fairness changes, is presented in detail in the next chapter.
Chapter 4

ENHANCING THE PERCEIVED FAIRNESS OF ALGORITHMIC DECISION-MAKING VIA LESS RESTRICTIVE EXPLANATIONS

This chapter presents the second empirical study, which focused on examining the impact of explanations’ restrictiveness on organizational decision-makers’ perception of the fairness of algorithmic decision-making. The first section of the chapter introduces the research question related to experiment 2 and the importance of it. Then, I delve into the tension between decisional guidance and restrictiveness, for a grounding that emphasizes the necessity of considering it if one wishes to understand the relationship between explanations and perceptions of the fairness of algorithmic decision-making. Section 3.3 describes the research model specific to this study. Following the explanation of the model and the hypotheses’ development, I finish setting the stage by discussing the research methods, together with details of the study design and the experiment-development process, before Section 4.5 lays out the results of the data analysis.

The results show that explanation restrictiveness influences all three dimensions of perceptions of fairness. Less restrictive counterfactual explanations increase the sense of procedural fairness. For perceptions of informational fairness, an interaction effect between less restrictive explanations of the model’s performance and less restrictive counterfactuals was visible; when both are less restrictive, there is less perceived informational fairness. The same interaction effect emerged for the sense of distributional fairness: jointly, less restrictive counterfactuals and model-performance explanations yielded a reduced sense of distributional fairness. These results are discussed in detail, and Section 4.6 rounds out the chapter with conclusions from the study.
4.1. Introduction to the Restrictiveness Study

Literature on human–computer interaction (HCI) attests that allowing the users to **explore** how analytics systems reach their decision is effective in producing greater understanding of the systems and the potential consequences that might arise (Bussone, Stumpf, & O’Sullivan, 2015; Kobsa, 1994). This exploration can be achieved through interactivity, a mechanism that gives the decision-maker greater exposure to the system and enables more in-depth thinking about the process of algorithmic decision-making as a whole, by providing broader-based insight on the context behind the system. Research has identified such benefits of interactive explanation as fuller understanding of the system and better task performance (Beaulieu & Jones, 1998; Johnson & Johnson, 1993; Li & Belkin, 2010; Wang & Benbasat, 2009). However, the impact of interactivity through less restrictive explanations with regard to decision-makers’ perceptions of fairness has not been explored in the context of algorithmic decision-making.

In the first experiment, providing counterfactual explanations restricted the interaction between the decision-maker and the analytics system by narrowing the room for interpretation of the decision made and hindering the decision-makers’ exercise of their personal judgment in the decision-making process. They did not question the decisions on behalf of the customers in light of how fair the decision might be from the customer’s angle. One can conclude, therefore, that explaining the decision by means of counterfactual reasoning and evaluations of system performance does not constitute a complete remedy – a way of rendering decision-makers skeptical about potentially unfair decisions by algorithms and, more importantly, of influencing their perceptions as to fairness. That was because the explanation given restricted the decision-makers’ judgment related to the performance of the system and the way it made the decision. This finding shed light on the importance of exploring the extent to which explanations should be guiding rather than restrictive, so as to create space for decision-makers to practice their own judgment rather than meekly follow the algorithmic decisions in lockstep (Afrashteh, Asadi Someh, & Davern, 2018; Edwards & Veale, 2017).

Accordingly, this study was designed to answer the question “How does restrictive explanation of algorithm’s decision influence the fairness perceived by organizational decision-makers?”
To address this, I developed a 2x2 between-subjects experiment in which the extent to which decision-makers can engage with the model-performance explanations and with the decision-making explanations are manipulated. I operationalized the less restrictive explanations via a more interactive explanations focused on the algorithmic decision-making business process, including the model’s development and the decision-making process. I theorized that these manipulations create the room required for decision-maker interpretation of the algorithmic decisions and consequently influences their perceived fairness.

As a backdrop to that operationalization, the next section provides theoretical background on decisional guidance and the tension between that guidance and restrictiveness. It also introduces prior work on interactivity. Then, I present my theory and research model, including specific hypotheses about the impact of less restrictive explanations on fairness perceptions. My presentation of the research method is followed by discussion of the data analysis. The chapter concludes with a summary of conclusions.

4.2. Background on Related Theory

4.2.1. Restrictiveness

Proceeding from the first experiment’s results, I posited that the explanations’ restrictiveness lies behind counterfactuals’ reduction in the room for interpretation. The literature, in which system restrictiveness, per Silver (1987), is a concept for the extent and manner of a system restricting the users’ decision-making processes to a particular subset of those possible, identifies two main sources of it: constraints that arise from the physical features and that arise from the interaction with the physical features (Lynch & Gomaa, 2003; Wheeler & Valacich, 1996). The former, physical restrictiveness, is the degree to which the functional capabilities of the system limit decision-makers’ use of it (Chu & Elam, 1990; Davern & Kamis, 2010; Tabatabaei, 2002). The sources of restriction to the interaction between the users and the system, in turn, can be divided into structural and process restrictiveness. Structural restrictiveness is related to the system-internal sources of restriction: features embedded in the systems that impose constraints on the interaction between the users and the system. Process restrictiveness,
Chapter 4 Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

on the other hand, involves the external interaction with the system and refers to the extent to which system-external factors impinge on users’ interaction with the system and the way they employ it (Lynch & Gomaa, 2003; Wheeler & Valacich, 1996).

My focus was on interaction-related restrictiveness especially, on account of the study’s context: it was intended to create room for interpretation for the decision-makers by communicating explanations. Since the evaluation metrics and evidence of reasoning explain the performance of the model and its decision from the standpoint of the internal workings of the system, I looked specifically at structural restrictiveness (sources of restriction internal to the system). This afforded an individual-behavior-linked view of the design aspect of the model and helped me consider the restrictiveness construct as a design feature through which space can be created. Additionally, focusing on external restrictiveness was out of scope for this study because it requires an understanding of several factors beyond the algorithmic model: the user’s personality, authority, role, etc. For this study, I looked at structural restrictiveness, understood as how the explanation limits decision-makers’ exercise of judgment about the decision made by the system.

Several DSS researchers have studied the role of system restrictiveness as a design characteristic with regard to users’ cognition and their performance (Davern & Kamis, 2010; Goodwin et al., 2011; Silver, 1987, 1988; Wang & Benbasat, 2009). However, research has not examined how restrictiveness may influence the cognitive operations of decision-makers asked to make their own judgments while guided by algorithms’ suggestions. The same is true for their perceptions of the algorithmic decisions’ fairness.

Again, this study followed on from the first empirical study. There, structural restrictiveness (counterfactual explanations) imposed limitations on how users interacted with the system and viewed its performance and output (i.e., it closed the door to interpretation).

4.2.2. Decisional Guidance

Given how explanations as inputs guide decision-makers toward applying their judgment about algorithms and algorithmic decisions, it is important to consider especially the attribute of explanations that is referred to as decisional guidance (see Silver, 1991). In the DSS literature, decisional guidance has been defined as how a DSS enlightens or sways its users “as they structure and execute their decision-making processes – that is,
as they choose among and use the system’s functional capabilities” (Silver, 1991, p. 107). There are two types of decisional guidance: suggestive and informative. The suggestive form provides judgment recommendations and direction for decision-makers on what information they may take as input for their judgment, and how, while informative decisional guidance provides relevant information to inform decision-makers’ judgment rather than make suggestions (Silver, 1991).

Applying this definition to the algorithmic decision-making context, one can define decisional guidance as how the explanation enlightens/sways the decision-makers as they employ algorithms in their decision-making – i.e., as it creates space for them to exercise their judgment as to the algorithms’ performance and decision-making logic. This study focused on the informative decisional guidance, since I sought explanations designed to create room for interpretation for the decision-makers.

4.2.3. Decisional Guidance or Restrictiveness

The DSS literature has always manifested a tension between decisional guidance and restrictiveness, each with its own influence on the behavior of a system’s users (Silver, 1987, 1991). Highlighting this tension, I explored the extent to which explanations have to be restrictive vs. guiding (in this context, informative) if they are to create the room for decision-makers to interpret the algorithmic decisions and exercise judgment. Creating this space, in turn, should affect their fairness perceptions. To this end, one approach is to offer an exploration opportunity whereby decision-makers can discover more about the model’s performance and the logic it employs to reach a decision. This opportunity should also facilitate decision-makers’ cognitive engagement with the explanations and their digestion of the information given. Therefore, the third study used interactivity, conceptualized as described below, to address the issue of restrictive explanations.

4.2.4. Interactivity

It is clear that interactivity is a cognitively relevant feature of the system design, and scholars have conceptualized and defined it in several ways (Bonito, Burgoon, & Bengtsson, 1999). One of the more common definitions employed in human–computer
interaction literature characterizes it as the information flow between the computer and its user, with facets encompassing the task environment, system environment, interface area, input, output, and feedback (Tripathi, 2011). My study focused on the input flow, the information provided to the user for accomplishing the task (decision-making in this case).

The HCI literature suggests that providing interactive explanation improves user understanding and comprehension of the system’s behavior. The reason is that it enables the users’ exploration of the complex black-box behavior by letting them “play with” those features within their reach (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018; Cheng et al., 2019). However, it has been argued also that interactive explanations are more time-consuming (Cheng et al., 2019). In addition, IS literature suggests that, though system interactivity can lead to better decisions by human decision-makers, their decisions might not be efficient – this setting can prolong the time required for the decision-makers to grow engaged with the decision-making process cognitively. At the same time, it can volatilize system-users’ performance (Davern, Shaft, & Te’eni, 2012).

In this study, interactivity was conceptualized as the explanations’ ability to engage the decision-maker cognitively with the decision-making process in pursuit of more informed decisions.

4.3. The Research Model

Figure 4.1 presents this study’s overall research model. With my model and the associated theory, I sought to explain the impact of communicating less restrictive explanations with the organizational decision-makers on their perception of the fairness of decisions made by algorithms. To measure perceptions of fairness more accurately, one can break it down into dimensions (Binns et al., 2018). This break-down is discussed next.

Another key construct in the model is interactivity in communication (used here as a manipulation of the explanations’ restrictiveness, as mentioned in the previous section, in line with the cognitive research literature: Davern et al., 2012; Dickmeyer, 1983; Kottemann, Davis, & Remus, 1994; McIntyre, 1982). Alongside these elements, I considered the tension between decisional guidance and restrictiveness in communicating more interactive explanations (Benbya et al., 2021; Silver, 1987, 1991). The constructs’ definitions are discussed next.
4.3.1. Decision Outcomes

The outcomes of interest were the perceived fairness of the model’s performance and that of the decision made. To assess how well the decision-makers could distinguish between fairness and other, related perceptions and to help them distinguish among these, I measured their perceptions of satisfaction and reasonability also.

For this study, the fairness variable was examined in terms of three dimensions: informational, procedural, and distributional fairness. In summary, informational fairness refers to the extent to which the decision-makers consider the information they receive through the explanation to match what they need, procedural fairness is related to the decision-making process and the logic behind it, and distributional fairness is connected purely to the decision itself and refers to the extent to which decision-makers regard the decision as equitable for the person it concerns when that person’s circumstances are taken into account. The fourth aspect of fairness identified in the literature and listed in Chapter 2, interactional fairness, was not considered in this study. It refers to the extent to which the decision-receivers are treated respectfully by the decision-makers (Binns et al., 2018), and the experiment scenario involved no interaction between decision-makers and customers that could be evaluated for its fairness.
4.3.2. Interactive Model-Performance Explanations

The model-performance explanations from the first experiment were retained so that the impact of communicating them with greater interactivity could be explored. One part of the business process connected with algorithmic decision-making for addressing a given problem (here, handling requests for a higher credit limit) is model development, which data scientists must work through to select the most appropriate model from among several candidate models. Organizations have found it challenging to select appropriate models. Since there is a wide range of algorithms at their disposal, the data scientists and managers involved often find it difficult to choose the most suitable one. Therefore, one possible application of model-performance explanations is to assist with determining which model is appropriate. In addition, interactive model-performance explanations can spark greater cognitive engagement by decision-makers and can provide grounding for more informed decisions.

One way of operationalizing interactive model-performance explanations with regard to the model-development process is to give the decision-makers information about the performance of alternative models (Leyton-Brown, Nudelman, Andrew, McFadden, & Shoham, 2003; Rice, 1976). This information enables the decision-makers to compare the performance of the selected model with that of other, competing models and recognize that selecting the most appropriate model always involves a tradeoff. That awareness may make an impact on how decision-makers view the fairness of the model chosen and its decisions. Moreover, this tradeoff implies that there is not always a model that is necessarily best, with the highest level of performance for solving a particular business problem – the act of choosing the most appropriate model for a general case entails uncertainty for divergent problems. By accentuating that model selection is a judgment and choice, less restrictive and more interactive model-performance explanations give the decision-maker the space and freedom to form alternative interpretations of the outputs.

Therefore, I expected to find that presenting details of alternative models’ performance would produce different perceptions of the selected one’s procedural, informational, and distributional fairness. Decision-makers who are informed about the performance of alternative models might perceive selection of the most appropriate model for decision-making – and consequently the model’s output – as fuzzy. The procedural and distributional fairness experienced might decrease. However, since interactive
model-performance explanations give the decision-makers more of the comparative information required, their sense of informational fairness might end up greater. Consequently, I hypothesized thus:

**H1**: Decision-makers provided with interactive model-performance explanations are more likely to

- a. Perceive the level of procedural fairness to be lower
- b. Perceive the level of informational fairness to be higher
- c. Perceive the level of distributional fairness to be lower

than decision-makers provided with static model-performance explanations

It should be noted that whether decision-makers would change the decision was not within this study’s scope. The focus was on exploring perceptions of the decision offered.

### 4.3.3. Interactive Counterfactual Explanations

After the choice of model comes the second stage in the business process for algorithmic decision-making. With less restrictive and more interactive decision-making explanations, I aimed to expose the logic behind the decision more extensively, by providing information on the most important factors on which the decision is based, together with the decision boundaries. I retained counterfactual explanations from the previous study, with the goal of exploring the effect of their restrictiveness on perceived fairness.

One way to operationalize interactive counterfactuals is through an opportunity for the decision-makers to manipulate the representations of the explanations. This includes the sorting order for details of the various decisions made about customers with a profile similar to that of the customer in question, on whose attributes the current decision was based. This interactive information enables the decision-makers to compare the decisions produced for roughly similar customers, given their profiles and the decision boundaries, and recognize through these whether there are any contradictions. Such information would be expected to influence their perceptions of fairness. Table 4.1 exemplifies how a decision-maker could compare between a task case involving denial and cases of granting a credit increase.
Table 4.1: Example presentation of comparative data

<table>
<thead>
<tr>
<th></th>
<th>Monthly salary</th>
<th>Monthly expenses</th>
<th>Credit limit requested</th>
<th>Credit history (days overdue in the past year)</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 3</td>
<td>$3,600</td>
<td>$3,250</td>
<td>$5,500</td>
<td>1</td>
<td>Approve</td>
</tr>
<tr>
<td>Current customer</td>
<td>$4,100</td>
<td>$3,120</td>
<td>$6,000</td>
<td>13</td>
<td>Deny</td>
</tr>
<tr>
<td>Customer 1</td>
<td>$4,700</td>
<td>$3,500</td>
<td>$6,000</td>
<td>0</td>
<td>Approve</td>
</tr>
<tr>
<td>Customer 4</td>
<td>$3,900</td>
<td>$3,200</td>
<td>$6,500</td>
<td>5</td>
<td>Deny</td>
</tr>
<tr>
<td>Customer 2</td>
<td>$4,500</td>
<td>$3,350</td>
<td>$7,000</td>
<td>7</td>
<td>Deny</td>
</tr>
</tbody>
</table>

A decision-maker presented with this table can see that Customer 1 received approval even though his monthly salary is $500 less and his expenses $130 more than the case customer’s. The credit increase requested by Customer 1, however, is $500 less, and only one day of delay has been recorded for him. This comparison creates room for interpretation by the decision-maker and sheds light on the fuzziness of the decision made. The fuzziness is most salient in relation to edge cases (here, situations in which a given decision is somewhat ambiguous and minor changes in conditions would have led to a different recommendation). Its fuzzy nature is crystallized even more via the interactivity provided for the decision-makers: adjusting the presentation order of the various customer profiles aids in comparing decisions. I developed the following hypothesis specific to interactivity as follows:

**H2:** Decision-makers provided with interactive counterfactual explanations are more likely to

a. Perceive the level of procedural fairness to be lower

b. Perceive the level of informational fairness to be higher

c. Perceive the level of distributional fairness to be lower

than decision-makers provided only with static counterfactuals
4.4. The Research Method

4.4.1. Participants

As those in the previous experiment were, participants were given information about a credit application and asked to determine, in light of the decision made by the algorithm, whether or not the customer’s request for a higher credit limit should be approved. Again, all subjects were native English-speakers over the age of 25 who were registered with the Prolific.co online subject-pool platform. For greater external validity of the project, the participants were pre-filtered for only Prolific users who had earned at least an undergraduate degree. To guarantee sufficient power to test the hypotheses, I recruited 260 individuals (i.e., 65 per cell in a 2×2 design).

Data for 25 participants were filtered out from analysis because of either an unreasonably short time for completion of the study (7 subjects) or incorrect understanding of the experiment’s manipulation (indicated by giving incorrect answers to the two manipulation-check questions) (18 subjects). Of the 235 remaining participants, 144 were male and 91 female. As for education, 144 had an undergraduate degree, 84 a sub-doctoral graduate degree, and seven a PhD. Most participants were in their mid-thirties. The average participant was proficient in data analytics, with subjects’ average experience in credit-decision-making coming to 2.9 years. Finally, before analysis, the participants were filtered for management experience in the finance and insurance industry. The participants’ background data can be found in Table 4.2.

---

1 Analysis was performed also without removal of the participants who did not pass the manipulation checks. Although its results were not statistically significant, there were visible effects in the same direction as without their inclusion.
4.4.2. The Task and Procedures in the Experiment

A 2x2 between-subjects experiment design was employed, contrasting interactive vs. static model-performance explanations and interactive vs. static decision-making explanations (i.e., counterfactuals). The dependent variables were participants’ perceptions of satisfaction, reasonability, and fairness related to the decision.

In this study too, a 16-subject pilot test was conducted prior to the experiment, to make sure the process was understandable and the case scenario appropriate. This also aided in judging completion times: the average time for completing the decision-making task was roughly 10 minutes. The participants’ feedback proved valuable for the actual experiment.

Several personality variables that are potential confounding factors were addressed via control variables. I handled other generalizability issues by means of a scenario-based method. As noted in the previous chapter, research demonstrates that how people act in

<table>
<thead>
<tr>
<th>Table 4.2: Participant characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td><strong>Age</strong></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Level of education</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Data-analytics experience</strong></td>
</tr>
<tr>
<td><strong>Experience in making credit decisions</strong></td>
</tr>
</tbody>
</table>

*Note1: Sample size= 235. No missing data.*

*Note2: Data analytics experience and Credit decision making experience are 7-point scales with anchors 1=none and 7= substantial*
scenario-based experiments is consistent with their actions as they go about their day-to-day life (Woods et al., 2006). For this experiment too, I used a Web-based environment, utilizing the Qualtrics platform with participants sourced from Prolific.

As the first experiment did, this one entailed a simulated credit-limit increase decision by a manager at a financial institution that uses an analytics system to make a recommendation as to whether an increase should be granted. The purpose of the experiment design and the manipulations of exposure to model-development and to decision-making-logic information was to 1) prompt the participants in various ways to recognize that the hypothetical analytics system does not necessarily offer perfect performance and 2) explore whether interactivity produces awareness of possibly unintended consequences of algorithmic decision-making, thereby ultimately influencing the decision-maker’s perceptions of the decision. The scenario was designed to articulate an edge case: a customer profile was generated that lies close to the decision boundaries and hence might be placed in the wrong class by the algorithmic classifier. The human decision-maker, sometimes aided by the interactive facility, could judge whether the analytics system’s performance and the decision might lead to unfair consequences for the customer. In that event, the participant could override the decision made by algorithm, superseding it with a decision that he or she viewed as fairer.

4.4.3. Materials for the Experiment

The participants were provided with four scenarios within which the interactivity of the model-performance explanations and the counterfactual explanations were manipulated.

<table>
<thead>
<tr>
<th>Model performance explanations</th>
<th>Counterfactual explanations</th>
<th>Static</th>
<th>Interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>Static</td>
<td>54</td>
<td>59</td>
</tr>
<tr>
<td>Static</td>
<td>Interactive</td>
<td>63</td>
<td>59</td>
</tr>
<tr>
<td>Interactive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

Table 4.3 shows the number of participants exposed to each set of conditions. The dependent variables assessed in this study were perceptions of satisfaction, reasonability, and fairness. These variables were measured for two distinct targets: the performance of the system, then the decision by the system.

All dependent variables were measured on a six-point semantic differential scale, again to guarantee that the participants were not biased toward a given direction. There was also direct measurement via decision choice: whether or not the participants chose to overrule a denial decision recommended by the analytics. Binary coding was applied for the decision choice: the code “1” was assigned for when participants overrode the decision in order to grant approval, and the code “0” was assigned if they concurred with the analytics decision.

Control Variables

“Individuals’ tendency for engaging in thinking and enjoying [the fruits] from that” is captured by the “need for cognition” variable (Cacioppo & Petty, 1982, p. 118). This variable may be connected with the extent to which the participants are willing to become cognitively engaged with the explanations given and, consequently, experience effects on their fairness perceptions. Therefore, I controlled for the potential influence of this variable (NFC) on the decision-makers’ perceptions of fairness, to obtain more accurate results with regard to how restrictiveness of the explanations can influence decision-makers’ cognitive engagement and its impact on the fairness perceived.

Faith in general technology was taken into consideration as another control variable, since I found it to exert a significant influence in the first experiment.

I conducted exploratory factor analysis to assess the discriminant and convergent validity of the control variables.

Table 4.4 presents the results. Two components were extracted, with a cumulative percentage of 62%. In other words, the two factors extracted together explain 62% of the variance. The rotated component matrix shows that both “faith in general technology” and “need for cognition” have strong factor loadings and all items related to each of these variables are loaded together without displaying cross-loadings.
To assess the measurement properties of the instruments for the control variables and formulate factor scores, I performed confirmatory factor analysis. I began by assessing the internal consistency of the instruments. For FGT, the Cronbach’s alpha was 0.84, which put it above the threshold and, therefore, demonstrated internal consistency (Barclay et al., 1995; Fornell & Larcker, 1981). Then, the loading of each instrument item on the corresponding factor was assessed for FGT. This enabled me to check the reliability of the items. The loadings of all items were above 0.7 and hence met the criteria. Therefore, they are considered excellent, and the factor-analysis results can be deemed satisfactory (Tabachnick & Fidell, 2001). The loadings are presented in Table 4.5. Faith in technology, defined as the extent to which users assume that technologies are usually reliable and consistent, can be important: as noted in Chapter 3, some people put blind faith in algorithms and refuse to question their performance. Since people differ in their level of trust in algorithms, their perceptions of an algorithmic system and the decision made by analytics are bound to vary accordingly.

Table 4.4: Rotated component matrix

<table>
<thead>
<tr>
<th>Variable’s name</th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFC 1</td>
<td>0.839</td>
<td>-0.064</td>
</tr>
<tr>
<td>NFC 2</td>
<td>0.862</td>
<td>-0.007</td>
</tr>
<tr>
<td>NFC 3</td>
<td>0.772</td>
<td>0.041</td>
</tr>
<tr>
<td>NFC 4</td>
<td>0.760</td>
<td>0.122</td>
</tr>
<tr>
<td>NFC 5</td>
<td>0.583</td>
<td>-0.055</td>
</tr>
<tr>
<td>FGT 1</td>
<td>0.010</td>
<td>0.805</td>
</tr>
<tr>
<td>FGT 2</td>
<td>-0.030</td>
<td>0.856</td>
</tr>
<tr>
<td>FGT 3</td>
<td>0.001</td>
<td>0.857</td>
</tr>
<tr>
<td>FGT 4</td>
<td>0.013</td>
<td>0.749</td>
</tr>
</tbody>
</table>

Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

Table 4.5: Factor-analysis results for general faith in technology

<table>
<thead>
<tr>
<th>Variable’s name</th>
<th>Measurement item</th>
<th>Std. loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGT 1</td>
<td>“I believe that most algorithmic decision-making systems are effective at what they are designed to do.”</td>
<td>0.779</td>
</tr>
<tr>
<td>FGT 2</td>
<td>“A large majority of algorithmic decision-making systems are excellent.”</td>
<td>0.823</td>
</tr>
<tr>
<td>FGT 3</td>
<td>“Most algorithmic decision-making systems have the features needed for their domain.”</td>
<td>0.830</td>
</tr>
<tr>
<td>FGT 4</td>
<td>“I think most algorithmic decision-making systems enable me to do what I need to do.”</td>
<td>0.738</td>
</tr>
</tbody>
</table>

The SPSS statistics package was used for the factor analysis. Extraction method: maximum-likelihood, with no rotation since only one factor was explored.

The NFC variable too was assessed for its value as a control variable. Confirmatory factor analysis was performed for this variable likewise, to assess the measurement properties of the control-variable instruments and arrive at factor scores. The Cronbach’s alpha was 0.75. Therefore, NFC meets the criteria, so it was incorporated as a control variable. As Table 4.6 shows, I examined the loading of each instrument item on its corresponding factor, to check the reliability of the items. The items’ loadings met the standard criteria: they were all above 0.7.

Table 4.6: Factor-analysis results for need for cognition

<table>
<thead>
<tr>
<th>Variable’s name</th>
<th>Measurement item</th>
<th>Std. loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFC 1</td>
<td>“I don’t like to do a lot of thinking.”</td>
<td>0.870</td>
</tr>
<tr>
<td>NFC 2</td>
<td>“I try to avoid situations that require thinking in depth about something.”</td>
<td>0.790</td>
</tr>
<tr>
<td>NFC 3</td>
<td>“I prefer to do something that challenges my thinking abilities rather than something that doesn’t require much thought.”</td>
<td>0.714</td>
</tr>
<tr>
<td>NFC 4</td>
<td>“I prefer complex problems to simple problems.”</td>
<td>0.703</td>
</tr>
<tr>
<td>NFC 5</td>
<td>“Thinking hard and for a long time about something gives me little satisfaction.”</td>
<td>0.769</td>
</tr>
</tbody>
</table>

SPSS was employed for the factor analysis. Extraction method: maximum-likelihood, with no rotation since only one factor was explored.
Additionally, I applied data analysis for thorough examination of the justification stated by the participants, to uncover the main reason behind the final decision they made when overruling the one produced through analytics. The results show that the contextual information describing the health circumstances of a specific customer influenced the participants’ adoption of the customer’s perspective, and it also affected their perceptions of satisfaction, reasonability, and fairness related to the decision produced by means of analytics. This impact may be connected with how crucial people deem health-related issues to be. Therefore, a context-awareness variable was introduced for this study too. It was defined in the analysis and used as a covariate for statistically removing the impact of the context-specific information about the customer on the dependent variables. The instances of considering the contextual information on the reason behind the customer’s late payment were assigned the code “1” to represent the participant’s context-awareness, while cases of not taking the customer-context information into account were given the code “0.”

4.5. Data Analysis and Results

4.5.1. Descriptive Statistics

Pertinent descriptive statistics for the dependent variables related to fairness and for the participants’ decisions relative to those recommended are provided in Error! Reference source not found. and Table 4.8, respectively. The tables present the data after pre-analysis screening. In all, 235 participants completed the entire experiment, with all response sets being subjected to a manipulation check for gauging the extent of the participant’s awareness of the interactive model explanations and the interactive decision-making explanations of the material presented to them. I performed 2×2 factorial analysis of covariance to examine the effects of less restrictive model performance explanations and less restrictive counterfactual explanations on perceived satisfaction, reasonability, and fairness related to both the system’s performance and the algorithm’s decision, once I had accounted for the effects of covariates: faith in general technology, need for cognition, and context-awareness.
Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

### Table 4.7: Descriptive statistics for the dependent variables

<table>
<thead>
<tr>
<th>Construct</th>
<th>Mean</th>
<th>Median</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction (with the system’s performance)</td>
<td>4.20</td>
<td>5</td>
<td>1.347</td>
</tr>
<tr>
<td>Reasonability (of the system’s performance)</td>
<td>4.23</td>
<td>4</td>
<td>1.278</td>
</tr>
<tr>
<td>Fairness (of the system’s performance)</td>
<td>3.97</td>
<td>4</td>
<td>1.333</td>
</tr>
<tr>
<td>Satisfaction (with the algorithm’s decision)</td>
<td>4.25</td>
<td>5</td>
<td>1.314</td>
</tr>
<tr>
<td>Reasonability (of the algorithm’s decision)</td>
<td>4.28</td>
<td>5</td>
<td>1.371</td>
</tr>
<tr>
<td>Informational fairness (of the algorithm’s decision)</td>
<td>5.79</td>
<td>6</td>
<td>0.902</td>
</tr>
<tr>
<td>Procedural fairness (of the algorithm’s decision)</td>
<td>5.37</td>
<td>6</td>
<td>1.215</td>
</tr>
<tr>
<td>Distributional fairness (of the algorithm’s decision)</td>
<td>4.05</td>
<td>4</td>
<td>1.705</td>
</tr>
</tbody>
</table>

All measurements are on a six-point semantic differential scale with anchors 1 = negative pole and 6 = positive pole.

### Table 4.8: The decisions’ choice frequencies

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Decision to overrule the system</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deny (coded as 0)</td>
<td>Approve (coded as 1)</td>
</tr>
<tr>
<td>Interactive model-performance explanations × interactive counterfactuals</td>
<td>21 (35.6%)</td>
<td>38 (64.40%)</td>
</tr>
<tr>
<td>Interactive model-performance explanations × static counterfactuals</td>
<td>20 (31.74%)</td>
<td>43 (68.25%)</td>
</tr>
<tr>
<td>Static model-performance explanations × interactive counterfactuals</td>
<td>27 (45.76%)</td>
<td>32 (54.23%)</td>
</tr>
<tr>
<td>Static model-performance explanations × static counterfactuals</td>
<td>18 (33.33%)</td>
<td>36 (66.66%)</td>
</tr>
<tr>
<td>Total</td>
<td>86 (36.59%)</td>
<td>149 (63.40%)</td>
</tr>
</tbody>
</table>

Though the final decision by the organizational decision-maker is not the outcome of central interest, the frequencies of contrary decisions in each condition are presented in Table 4.8. There is a clear pattern: more of the participants (68.25%) overrode the denial decision to approve the request when provided with an interactive model performance and static counterfactuals in combination, whereas 63.40% of the participants given a matrix accompanied by counterfactual reasoning approved the non-recommended credit-limit increase.
4.5.2. Manipulation Checks

Of the four treatments, 74.4% of the participants, on average, recalled the alternative-models manipulation correctly and 69.8%, on average, recalled the manipulation of the interactivity correctly. Overall, these averages are high enough for concluding that the independent variables were designed well and acceptably recalled.

4.5.3. Correlations

Per the correlation matrix provided in Table 4.9 some interesting patterns are visible. The interactive model-performance explanations had a significant positive impact on the reasonability and fairness perceived with regard to the system’s performance, at $p = 0.05$ level. Secondly, prior experience with data analytics correlated positively with perceiving the decision to be satisfactory, as did context-awareness and overturning the recommendation. Perhaps unsurprisingly, the more satisfying and reasonable the decision was perceived to be, the fewer decisions were flipped to approval. Equivalent correlation can be seen between decision choice and informational, procedural, and distributional fairness. The metric for faith in technology was considered for applicability as a covariate. Since the correlations showed that faith in technology significantly affected all the dependent variables (the satisfaction, reasonability, and fairness perceptions for both system performance and the decision by the analytics), with $p = 0.001$, it was deemed a covariate for all further analysis. Since a meaningful significant negative correlation emerged between context-awareness and decision-linked perceived satisfaction, reasonability, and procedural and distributional fairness, context-awareness too was taken as a covariate in analysis of the dependent variables.
Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

4.5.4. ANCOVA Results

Analysis of covariance was conducted to examine the effects of the interactive model-performance explanations and the interactive counterfactual explanations on the perceptions of satisfaction, reasonability, and fairness related to both the system’s performance and the decision made by the systems. The results are reported in Table 4.10 shows that the perceived reasonability of the system’s performance varied between the condition with the evaluation metrics for alternative models’ performance and that without them. This effect is significant ($p = 0.032$). Nearly 10% of the difference visible

Table 4.9: Correlation matrix for the variables

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive explanation for the model’s performance</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>0.013</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived reasonability (of system’s performance)</td>
<td>0.150</td>
<td>0.095</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived satisfaction (with system’s performance)</td>
<td>0.1</td>
<td>0.115</td>
<td>0.804</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived fairness (of system’s performance)</td>
<td>0.181</td>
<td>0.075</td>
<td>0.748</td>
<td>0.743</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived reasonability (of decision by analytics)</td>
<td>-0.032</td>
<td>0.120</td>
<td>0.316</td>
<td>0.280</td>
<td>0.249</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived satisfaction (with decision by analytics)</td>
<td>0.014</td>
<td>0.097</td>
<td>0.240</td>
<td>0.279</td>
<td>0.274</td>
<td>0.805</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived informational fairness</td>
<td>0.045</td>
<td>-0.033</td>
<td>0.082</td>
<td>0.047</td>
<td>0.070</td>
<td>0.290</td>
<td>0.243</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived procedural fairness</td>
<td>-0.015</td>
<td>0.117</td>
<td>0.303</td>
<td>0.274</td>
<td>0.297</td>
<td>0.694</td>
<td>0.681</td>
<td>0.389</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived distributional fairness</td>
<td>0.025</td>
<td>-0.022</td>
<td>0.187</td>
<td>0.158</td>
<td>0.225</td>
<td>0.663</td>
<td>0.690</td>
<td>0.174</td>
<td>0.671</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision choice</td>
<td>0.61</td>
<td>-0.086</td>
<td>-0.005</td>
<td>0.046</td>
<td>-0.001</td>
<td>-0.471</td>
<td>-0.522</td>
<td>-0.151</td>
<td>-0.398</td>
<td>-0.496</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faith in technology</td>
<td>0.024</td>
<td>0.020</td>
<td>0.308</td>
<td>0.223</td>
<td>0.362</td>
<td>0.266</td>
<td>0.274</td>
<td>0.326</td>
<td>0.339</td>
<td>0.242</td>
<td>-0.082</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Awareness of context</td>
<td>0.042</td>
<td>-0.044</td>
<td>0.059</td>
<td>0.066</td>
<td>0.048</td>
<td>-0.273</td>
<td>-0.256</td>
<td>-0.003</td>
<td>-0.149</td>
<td>-0.373</td>
<td>0.519</td>
<td>0.069</td>
<td>1</td>
</tr>
</tbody>
</table>

Levels of p-values: * = $p \leq 0.05$, ** = $p \leq 0.01$, and *** = $p \leq 0.001$. 
in the dependent variable is explained by this analysis (the adjusted $R^2$ is 0.097). The interactive model-performance explanations had a significant positive effect on the perceived reasonability of system performance: providing details of alternative models’ performance increased the system-performance reasonability perceived by the subjects.

Table 4.11 shows that perceptions of satisfaction with the system’s performance did not vary between the interactive and the static model-performance explanations.

The figures in Table 4.12, in turn, suggest that the perceived fairness of the algorithmic model’s performance differed between the condition with interactive and that with static model-performance explanations. This effect is significant, with $p = 0.008$. Around 14% of the difference detected in the dependent variable is explained by this analysis, with an adjusted $R^2$ of 0.141. The presence of evaluation metrics for alternative models had a significant positive impact on the perceived fairness of the case system’s performance. Their presence led to greater perceived fairness of the system’s performance.

Table 4.13 shows that no significant effect of the interactive model-performance explanations and interactive counterfactuals was visible with regard to how reasonable the decisions made by the algorithm were deemed to be. Also attesting to a lack of statistically significant effect,

Table 4.14 addresses the interactive model-performance explanations and interactive counterfactuals with regard to perceptions of the satisfaction of the decision yielded by the scenario’s algorithm.

Table 4.15 represents the interaction effect exerted by the interactive model-performance explanations and the interactive counterfactuals jointly on the perceived informational fairness connected with the decision by the algorithm. This effect implies that the impact of the interactive performance explanations on the decision’s perceived informational fairness depends on the interactivity of the counterfactuals supplied. More specifically, when interactive model-performance explanations were provided, the informational fairness reported was at its lowest in the treatment with the interactive counterfactual explanations. This effect is significant, with $p = 0.015$. Approximately 12% of the difference in the relevant dependent variable is explained by this analysis (adjusted $R^2 = 0.123$). In the analysis of informational-fairness perceptions related to the decisions, only faith in technology was considered as a covariate, not
context-awareness. This is because I observed no initial correlation between the latter variable and the perceived informational fairness.

Table 4.16 illustrates the impact of interactive counterfactuals on the perceived procedural fairness of the algorithm’s decisions. This analysis shows divergence in the dependent variable between when the counterfactuals were static and when they were interactive. The effect is significant ($p = 0.059$). This analysis explains 15.5% of the difference evident in the dependent variable (i.e., adjusted $R^2 = 0.155$). The interactive counterfactuals had a significant positive impact on perceived procedural fairness. Their provision led to greater perceived fairness of the process within which the decision is made by algorithms.

The final table in the set, Table 4.17, presents an interaction effect of the interactive model-performance explanations and the interactive counterfactuals on the perceptions of distributional fairness related to the algorithmic decision. This effect implies that the interactive performance explanations’ impact on perceived distributional fairness is contingent on the interactivity of the counterfactuals supplied: when interactive explanations were provided, the ratings for distributional fairness were at their lowest in the treatment with interactive counterfactual explanations. This effect is significant, with $p = 0.049$. Approximately 10% of the difference detected in the dependent variable is explained by this analysis (adjusted $R^2 = 0.098$). In the analysis of the distributional fairness perceived in relation to the decisions, both faith in technology and need for cognition were accounted for as covariates. While an initial correlation was seen between context-awareness and perceived distributional fairness, its treatment as a covariate had no significant impact on the analysis; therefore, it was not regarded as one.
Chapter 4 Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

Table 4.10: ANCOVA results with perceived reasonability of the system’s performance as the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>6.869</td>
<td>6.869</td>
<td>4.657</td>
<td>0.032</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>23.851</td>
<td>23.851</td>
<td>16.170</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>3.068</td>
<td>3.068</td>
<td>2.080</td>
<td>0.151</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>0.489</td>
<td>0.489</td>
<td>0.332</td>
<td>0.565</td>
</tr>
<tr>
<td>Error</td>
<td>230</td>
<td>286.153</td>
<td>1.475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>3,886.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.097$.

Table 4.11: Results of ANCOVA where the dependent variable is perceived satisfaction with the system’s performance

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>3.434</td>
<td>3.434</td>
<td>2.006</td>
<td>0.158</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>11.804</td>
<td>11.804</td>
<td>6.894</td>
<td>0.009</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>3.068</td>
<td>3.068</td>
<td>2.080</td>
<td>0.151</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>8.133</td>
<td>8.133</td>
<td>4.750</td>
<td>0.030</td>
</tr>
<tr>
<td>Error</td>
<td>230</td>
<td>332.151</td>
<td>1.712</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>3,886.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.057$. 

97
Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

Table 4.12: Results of ANCOVA where the perceived fairness of system performance is the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>10.832</td>
<td>10.832</td>
<td>7.097</td>
<td>0.008</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>38.218</td>
<td>38.218</td>
<td>25.038</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>2.797</td>
<td>2.797</td>
<td>1.832</td>
<td>0.177</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>0.185</td>
<td>0.185</td>
<td>0.121</td>
<td>0.728</td>
</tr>
<tr>
<td>Error</td>
<td>230</td>
<td>296.117</td>
<td>1.526</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>3,488.000</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.141$.

Table 4.13: ANCOVA results where the dependent variable is perceived reasonableness of the algorithm’s decisions

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.200</td>
<td>0.200</td>
<td>0.122</td>
<td>0.727</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>4.939</td>
<td>4.939</td>
<td>3.008</td>
<td>0.084</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>26.749</td>
<td>26.749</td>
<td>16.290</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>3.180</td>
<td>3.180</td>
<td>1.936</td>
<td>0.166</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>30.092</td>
<td>30.092</td>
<td>18.326</td>
<td>0.000</td>
</tr>
<tr>
<td>Interactive model-performance explanations * interactive counterfactuals</td>
<td>1</td>
<td>0.398</td>
<td>0.398</td>
<td>0.242</td>
<td>0.623</td>
</tr>
<tr>
<td>Error</td>
<td>230</td>
<td>315.275</td>
<td>1.642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>3,953.000</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.149$. 
Table 4.14: Results of ANCOVA where satisfaction perceptions related to the algorithm’s decisions are the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.181</td>
<td>0.181</td>
<td>0.120</td>
<td>0.729</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>3.405</td>
<td>3.405</td>
<td>2.265</td>
<td>0.134</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>24.420</td>
<td>24.420</td>
<td>16.909</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>13.282</td>
<td>13.282</td>
<td>8.835</td>
<td>0.003</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>23.630</td>
<td>23.630</td>
<td>15.718</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.176</td>
<td>0.176</td>
<td>0.117</td>
<td>0.732</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>0.350</td>
<td>0.350</td>
<td>0.492</td>
<td>0.484</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>18.337</td>
<td>18.337</td>
<td>25.793</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>0.767</td>
<td>0.767</td>
<td>1.079</td>
<td>0.300</td>
</tr>
<tr>
<td>Interactive model-performance explanations *</td>
<td>1</td>
<td>4.264</td>
<td>4.264</td>
<td>5.998</td>
<td>0.015</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error                                             194  286.653  1.503
Total                                             199  3,886.000

Adjusted $R^2 = 0.168$.

Table 4.15: ANCOVA results with the algorithmic decision’s perceived informational fairness as the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.204</td>
<td>0.204</td>
<td>0.287</td>
<td>0.593</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>0.350</td>
<td>0.350</td>
<td>0.492</td>
<td>0.484</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>18.337</td>
<td>18.337</td>
<td>25.793</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>0.767</td>
<td>0.767</td>
<td>1.079</td>
<td>0.300</td>
</tr>
<tr>
<td>Interactive model-performance explanations *</td>
<td>1</td>
<td>4.264</td>
<td>4.264</td>
<td>5.998</td>
<td>0.015</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error                                             194  136.501  0.711
Total                                             198  6,781.000

Adjusted $R^2 = 0.123$.
Chapter 4 Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

Table 4.16: ANCOVA results with the algorithmic decisions’ perceived procedural fairness as the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>$F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.022</td>
<td>0.022</td>
<td>0.018</td>
<td>0.894</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>4.492</td>
<td>4.492</td>
<td>3.615</td>
<td>0.059</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>31.189</td>
<td>31.189</td>
<td>25.103</td>
<td>0.000</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>7.140</td>
<td>7.140</td>
<td>5.746</td>
<td>0.017</td>
</tr>
<tr>
<td>Context-awareness (covariate)</td>
<td>1</td>
<td>6.456</td>
<td>6.456</td>
<td>5.196</td>
<td>0.024</td>
</tr>
<tr>
<td>Interactive model-performance explanations *</td>
<td>1</td>
<td>0.207</td>
<td>0.207</td>
<td>0.167</td>
<td>0.683</td>
</tr>
<tr>
<td>interactive counterfactuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>191</td>
<td>237.313</td>
<td>1.242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>5,975.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.155$.

Table 4.17: ANCOVA results where algorithmic-decision-related perceived distributional fairness is the dependent variable

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>$F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive model-performance explanations</td>
<td>1</td>
<td>0.017</td>
<td>0.017</td>
<td>0.007</td>
<td>0.935</td>
</tr>
<tr>
<td>Interactive counterfactuals</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.990</td>
</tr>
<tr>
<td>Faith in technology (covariate)</td>
<td>1</td>
<td>23.525</td>
<td>23.525</td>
<td>9.062</td>
<td>0.003</td>
</tr>
<tr>
<td>Need for cognition (covariate)</td>
<td>1</td>
<td>24.111</td>
<td>24.111</td>
<td>9.287</td>
<td>0.003</td>
</tr>
<tr>
<td>Interactive model-performance explanations *</td>
<td>1</td>
<td>9.650</td>
<td>9.650</td>
<td>3.717</td>
<td>0.049</td>
</tr>
<tr>
<td>interactive counterfactuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>187</td>
<td>485.462</td>
<td>2.596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>3,791.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adjusted $R^2 = 0.098$. 
4.6. Discussion

At this juncture, it is worth stressing the value of this study’s two-stage experiment design. This design was aimed at providing interactive explanations for two distinct phases in the business process of algorithmic decision-making: development of the model and its decision-making process. Since comparing candidate models’ performance and selecting one model from among many is a vital part of any model’s development, the information given to decision-makers followed suit. It was supplied in the first stage in the form of the interactive model-performance explanations. The results show that the interactive model-performance explanations increase decision-makers’ sense of the reasonability and fairness of the system’s performance. These results are consistent with prior reports positing that decision-makers, when provided with details of the alternative models’ performance, realize something about the tradeoffs expressed by the evaluation metrics, and their perceptions are influenced as they compare the models. Therefore, with regard to perceptions of system performance as a whole rather than the local, specific case, the results suggest that decision-makers tend to take the organization’s perspective. They deem the model in use to be fair and reasonable if it demonstrates a good balance of high accuracy and correct approvals against a low number of incorrect approvals. However, when it comes to the specific case in question, an interactive model-performance explanation facility does not display a significant impact on perceived reasonability. This is because alternative models’ evaluation metrics expose the performance of the system only from the angle of the development process, irrespective of the circumstances of the case. They are global, not decision-specific.

In the second stage, oriented toward explaining the decision-making process, the setting provided interactive counterfactuals. The results reveal that the decision-makers were more likely to take the customer’s perspective in this condition – they benefit from contextual information about the case. Moreover, as noted above, the interactive counterfactuals reveal more of the system’s process, in terms of the logic and deliberation employed to reach its final decision.

The results show that decision-makers’ perceptions as to the decision’s fairness are affected by provision of interactive counterfactuals. Measuring perceptions of fairness
Chapter 4  Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

on the informational, procedural, and distributional dimensions facilitated in-depth analysis of this phenomenon.

The information provided to explain the system’s performance and the decision made was perceived as less fair when both interactive model-performance explanations and interactive counterfactuals were supplied. In line with the definition of informational fairness, this result may be explained by the information-overload theory (Maes, 1995). In the experiment, participants were able to compare the system’s performance with that of other models and could also compare the decision made for the specific edge case in question with other decisions, made for cases involving similar circumstances. This could well constitute cognitive overload that left them overwhelmed and unable to process the information adequately.

In contrast, the results suggest that bringing interactivity into counterfactual explanation leads to greater perceived procedural fairness. This could stem from decision-makers’ provision with sufficient space to see how the decision was made for other edge-case customers. With transparency, they were enabled to compare the case in question to those of customers with similar profiles and roughly analogous repayment records. This comparison of the system’s predictions between the current customer and historical ones helped them to evaluate the logic of the decision-making process. When provided with only static counterfactuals, the decision-makers were unable to comprehend the decision-making process. These gave them a sense merely of the decision boundaries, which aid in recognizing respects in which the case customer could improve.

The results could be explained also by Toulmin’s model of argumentation theory, in the sense that, with the interactive counterfactuals, the decision-maker is furnished with all six components of a good and practical argument. The first four components proposed by Toulmin (the claim, data, warrant and backing) exist in static and interactive counterfactuals both (the decision corresponds to the claim in this model, the important features on which the decision is based correspond to the data element, the customer’s credit score constitutes the warrant, and the credit-rating threshold is the backing). Through interactivity, the two other components – grounds for perceiving the qualifier and rebuttal –are provided too. By comparing the current case with other customers’, the decision-maker perceives how certain the claim is and knows about the cases that could contest the evidence. Consequently, the entire argumentation entity that
Chapter 4 Enhancing the perceived fairness of algorithmic decision-making via less restrictive explanations

communicates the logic behind the decision-making is perceived as fairer (Janik, Rieke, & Toulmin, 1984).

The findings related to distributional fairness suggest that the level of fairness perceived in the decision decreases when interactive model-performance explanations are provided and the counterfactuals are interactive. The reason for this impact is that, as was expected, supplying material on the alternative models’ performance highlights the tradeoff in model evaluation and draws decision-makers’ attention to the fact that the algorithm’s performance is not necessarily perfect. With this awareness foregrounded, interactive counterfactuals may arouse skepticism about the fairness of a given decision relative to that visible in cases involving parallel profiles and histories.
Chapter 5

CONCLUSIONS AND DIRECTIONS FOR THE FUTURE

The final chapter integrates the results of the two studies presented in Chapter 3 and Chapter 4. The chapter also ties these in with the first study, the literature examination presented in Chapter 2.

It reviews the work in terms of the project’s aims and illuminates how the findings from this research have addressed the research questions. This synthesis is followed by a summary discussion of overall insights and the main contributions of the research for both academic scholars and IS practitioners. Finally, the chapter highlights the limitations of the doctoral project and their flip side, potential opportunities for future research.
5.1. Summary of Research Findings

The work behind the dissertation pursued understanding of the impact of explanations on the fairness perceived by organizational decision-makers in connection with algorithmic decision-making. With the studies now presented, I can return to the three research questions, posed in Chapter 1, and summarize the answers uncovered.

RQ1: What is the meaningful information that must be explained to the decision-makers at an organization when it employs algorithmic decisions?

I answered RQ1 by conducting comprehensive theoretical groundwork on explanations and the associated concepts that one must consider when looking at provision of explanations for the purpose of understanding their impact on the perceived fairness of the decision made. The first study pinpointed the importance of considering explanations as discourse for the purpose of influencing decision-makers’ perception of fairness. In line with this aim, key dimensions of explanations discourse was introduced. These consist of the explanations’ content types, explanation scope, the reasoning logic behind the explanations, and the strategies for communicating the explanations.

I posited that, within the context of algorithmic decision-making, to influence the organizational decision-makers’ sense of fairness, the “meaningful information” required of the explanations as discourse must encompass all four: suitable content, scope, reasoning logic, and a strategy for the communication. With regard to the content, it is important to explain the algorithm’s performance (via evaluation metrics) together with the logic behind how it reaches a decision (evidence of its reasoning). This is because these two facets cover the most important aspects of an algorithm: how well it performs and the process by which it arrives at its decision. With regard to scope, a holistic combination of local and global explanations can serve as a lens for understanding whether the decision produced is fair. Local explanations shed light on the individual decision at issue, while global ones illuminate the general logic of the decision-making process. As for reasoning logic, in turn, examining the contrast between factual and counterfactual explanations yields insight as to how views of fairness may be influenced when the manner of supporting the decision changes. Finally, the information needs to be communicated commensurately with the purpose
for its use if one wishes to achieve meaningfulness and consequently influence the fairness perceived by the decision-makers. Given the aim of this endeavor, the communication strategy of explanations must account for the perspectives of the core stakeholders in the decision-making process along with the extent to which explanations in the relevant context need to be guiding rather than restrictive.

All four dimensions are important in understanding of how explanations enhance algorithmic fairness. All of them are interdependent; depending on which strategy is taken for explanation communication, the content, the reasoning logic and the scope of explanations can be different through the discourse.

**RQ2: How does perspective-taking in the setting of explaining an algorithm’s decisions influence the fairness perceived by the organization’s decision-makers?**

Via my 2×2 experiment exploring the perceived-fairness impact of perspective-taking in explaining an algorithm’s performance and how it reaches its decision, I found that presenting the system’s error matrix – wherein an emphasis on false-negative rates guides the customer’s perspective – reduces the level of fairness, satisfaction, and reasonability perceived in the performance of the algorithm and the sense of the resulting decision’s reasonability. In contrast, presenting counterfactual explanations did not lead to the expected decrease in the perceived fairness of the algorithm’s decision: no significant influence was found. In fact, a pattern emerged in which counterfactuals had a positive impact on perceptions of the decision’s fairness. This offered evidence against the study’s second hypothesis.

**Table 5.1: Summary of the first empirical study’s results**

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Satisfaction</th>
<th>Reasonability</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System’s performance</strong></td>
<td>An error matrix’s presence decreases the sense of satisfaction</td>
<td>An error matrix’s presence decreases perceived reasonability</td>
<td>An error matrix’s presence decreases perceived fairness</td>
</tr>
<tr>
<td><strong>Algorithm’s decision</strong></td>
<td>Presenting a factual explanation decreases the sense of satisfaction</td>
<td>An error matrix’s presence decreases perceived reasonability</td>
<td>No significant effect was found</td>
</tr>
</tbody>
</table>


Chapter 5 Conclusions and directions for the future

Nonetheless, factual explanations did lower the satisfaction experienced with the decision made. Table 5.1 summarizes the results of the first study.

**RQ3: How does restrictive framing of the explanation for an algorithmic decision influence the fairness perceived by organizational decision-makers?**

The aforementioned counterintuitive finding from experiment 1, that counterfactuals increased perceived fairness, highlighted the role of explanations’ restrictiveness and the extent to which it can impede decision-makers from exercising their judgment about the decisions produced by algorithms. This is why I followed up with a 2×2 experiment to explore the impact of less restrictive explanation-framing on the perceptions of fairness. Experiment 2 used interactivity for operationalizing less restrictive explanations. For this study, I also broke fairness into three dimensions (informational, procedural, and distributional) to measure the impact more accurately, in finer granularity. I ascertained that interactive model-performance explanations and interactive counterfactual explanations each play an important role in the fairness and reasonability ultimately ascribed to the algorithm’s performance and the decision. Table 5.2 summarizes the final study’s results.

<table>
<thead>
<tr>
<th>Table 5.2: Summary of the second empirical study’s results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 2</strong></td>
</tr>
<tr>
<td><strong>System’s performance</strong></td>
</tr>
<tr>
<td>Satisfaction:</td>
</tr>
<tr>
<td>No significant effect was found</td>
</tr>
<tr>
<td>Reasonability:</td>
</tr>
<tr>
<td>Error-matrix interactivity increases perceived reasonability</td>
</tr>
<tr>
<td>Fairness:</td>
</tr>
<tr>
<td>Error-matrix interactivity increases perceived fairness</td>
</tr>
<tr>
<td><strong>Algorithm’s decision</strong></td>
</tr>
<tr>
<td>Satisfaction:</td>
</tr>
<tr>
<td>No significant effect was found</td>
</tr>
<tr>
<td>Reasonability:</td>
</tr>
<tr>
<td>No significant effect was found</td>
</tr>
<tr>
<td>Fairness:</td>
</tr>
<tr>
<td>Informational:</td>
</tr>
<tr>
<td>Error-matrix interactivity and presenting counterfactuals exerted an interaction effect on perceived fairness</td>
</tr>
<tr>
<td>Procedural:</td>
</tr>
<tr>
<td>Presenting interactive counterfactuals increased perceived procedural fairness</td>
</tr>
<tr>
<td>Distributional:</td>
</tr>
<tr>
<td>Joint presentation of an interactive error matrix and counterfactuals showed an interaction effect on perceived fairness</td>
</tr>
</tbody>
</table>

107
Chapter 5 Conclusions and directions for the future

5.2. Contributions and Implications for Scholarship

This research has implications for the literature on explanations, perspective-taking and decisional guidance as described below:

I extended the literature on explanations in following aspects:

First, I highlight the important role of explaining algorithms to the organizational decision-makers on their perceived fairness about the decision providing a new theoretical lens of discourse ethics. Bringing the discourse ethics theory into play, it opens the possibility for drawing theoretical insight from the established literature on organizational communication enabled by IT artefacts through highlighting the important role of explanations and their communication (Mingers & Walsham, 2010).

Second, I introduced the four dimensions of explanations which are required to be considered in providing “meaningful information” as explanations to make the most impact on the perception of fairness. These dimensions include the content types of explanations, the scope of explanations, the explanations’ reasoning logic and the explanations discourse. Introducing these dimensions extends the prior literature on explanations in intelligent systems and their associated constructs and taxonomies (Gregor & Benbasat, 1999) to the algorithmic decision-making literature.

Third, I provided important insights that contribute to IS literature in algorithmic decision-making when investigating how to address the gap between algorithmic models and decision-makers mental model (Martens & Provost, 2013). The findings of this research reflect on how explanations can lead to better understanding of algorithms by exploring the impact of different explanations with different framings to be discoursed with the organizational decision-makers on decision-makers’ perceptions about algorithmic decisions.

Eventually, the findings of this research can be incorporated with the other studies from social science, cognitive science and human-computer interaction with the aim of creating good explanations answering the call for contribution by Miller (2018).

My research contributes to perspective-taking literature as described below:
Chapter 5 Conclusions and directions for the future

To the best of my knowledge the perspective-taking theory introduced by Gasiorek and Ebesu Hubbard (2017) has not been extended to the algorithmic decision-making and explanations context. Thus, my study extends perspective taking theory to a new application context in information systems and contributes to the ongoing debate about the algorithmic decision-making fairness and its consequences for different stakeholders.

I contributed to existing literature on decisional guidance as described below:

To the best of my knowledge the notion of decisional guidance versus restrictiveness has not been conceptually specified or empirically examined in algorithmic decision-making and explanation context. In the Decision Support Systems (DSS) literature, M. S. Silver (1990) has illustrated how those systems can restrict or expand the decision-making process given the organizations requirements and goals. However, IS scholars have called for further research advancing the sociotechnical approach of IS in the explanations context, by drawing on the tension between automation and augmentation.

It is important to understand how organizations can employ algorithms and benefit from their high performance while being responsible for their potential consequences for individuals (Benbya et al., 2021). My study extends the decisional guidance literature within the context of algorithmic decision-making and explanations by highlighting the tension between guiding explanations versus restrictive explanations and how less restrictive explanations influence on the decision-makers’ perception of fairness. Moreover, the significant results found about the impact of explanations’ restrictiveness on the perception of fairness lead to a high-level understanding of how framing explanations in less restrictive way can play a role in making the decision-makers aware about potential unfair decisions which subsequently addresses the abovementioned scholars call for research.

5.3. Implications for Practice

My research has several practical implications. I present three important ones below.

This study provides insight surrounding how explanations can address the fairness issues caused by applying algorithms for decision-making. This insight points out that
providing explanations doesn’t necessarily lead to greater perception of fairness. It can be considered as a way towards transparency highlighting the potential fairness issues for the decision-makers. Moreover, it contributes to understanding of what could constitute meaningful information, in the GDPR’s terminology. This could inform the guidance for organizations that need to comply with such regulations yet find it difficult to get a clear idea of what constitutes explanations. Moreover, my work has yielded insight as to how an organization should frame and communicate its explanations to its decision-makers for an enhanced sense that the algorithms’ decisions are fair.

Secondly, it suggests design interventions for algorithm decision-making systems and the associated business process that show potential to enable a more balanced ethical outcome. The findings could inform efforts to reduce the reputational risk of using such systems while conferring richer appreciation for the quality of such systems’ decisions. The results may identify a means of empowering decision-makers to navigate appropriately between competing stakeholder perspectives. Thereby, all parties can be guaranteed a better outcome. Decision-maker feedback arising from situational awareness of different stakeholder perspectives holds potential to inform the refinement of practices related to machine learning and identification of biases in the underlying training data.

Finally, at the highest level, this research delivers important insights to organizations that invest in implementing or improving their data-management and data-governance programs. Data ethics is an important part of managing and governing data (DAMA International, 2017). Handling data in an ethical manner throughout the data life cycle is the key for all organizations that want to benefit from their data assets. As algorithms gain increasing prominence in organizations’ data analysis, which is a core element of the data life cycle, my work provides valuable guidance. Organizations must look at how they can employ explanations as means of transparency to address various unethical consequences that could arise from using algorithms in their decision-making processes.

5.4. The Project’s Limitations and Openings for Future Work

The first study did not examine all possible dimensions of explanations. Among the dimensions beyond the scope of this research were decision-maker characteristics such
as level of domain expertise, the explanations’ presentation format, and the media used for their provision (Gregor & Benbasat, 1999). Hence, the set of dimensions considered could be extended to include more factors that might contribute to an impact on perceived fairness. That impact could be achieved through considering them either when creating the content of the explanations or when framing and communicating them for the decision-makers.

The second study framed explanations through perspective-taking, with the decision-makers only being prompted to take the perspective of either customers or the organization. However, I did not measure whether the decision-makers actually adopted either of those perspectives. Although I used the context-awareness variable to understand whether the decision-makers took on the customer’s perspective or not, this could be measured more specifically and accurately in future. Doing so might help us understand more fully whether the framings of explanations developed truly represent the intended perspectives.

The study of explanations’ restrictiveness contributed to our understanding of specific sources of restrictiveness. It did not address other sources and how they influence decision-makers’ perceptions of fairness. As noted in the corresponding chapter, restrictiveness can stem from physical features or the interaction with those features. I focused on only the latter source. In addition, interaction-related restrictiveness can itself be broken down into sources that are internal vs. external to the explanations. The focus of the third study was on exploring the explanation-internal factors that restrict interaction between the explanations and the decision-makers. Research could be conducted that explores the impact of restrictiveness sources external to the explanations. These restrict the process of interaction by imposing any of several types of constraints on the decision-makers. The following sources are among those that further studies could address:

- The goal of explanation-provision: is it for improving the decisions and the algorithm’s performance, or is it just for making decision-makers aware of a potentially unfair decision, so that they can respond accordingly?
Chapter 5 Conclusions and directions for the future

- The type of the task that the decision-makers are asked to complete after provision of the explanation – for instance, just reviewing the decision or, in contrast, treating it as a recommendation and making the final decision on the basis of the explanations provided.

- The decision-makers’ role and whether they are authorized to overturn the algorithmically produced decision on the basis of their perceptions, along with the extent to which they are accountable for the consequences of the decisions.

In the Figure 5.1, dashed-red shapes encircle areas wherein future work could address further restrictiveness sources and their impact on perceived fairness.

**Figure 5.1:** Potential future work on sources of explanations’ restrictiveness

It is of note that the proposed future studies will also have some limitations. The main limitations would be the construct and external validity. Construct validity refers to “the possibility that the operations which are meant to represent a particular cause or effect construct can be construed in terms of more than one construct”(Campbell & Cook, 1979, p. 59). External validity relates to generalizing the results “to particular target persons, settings, and time” and “across types of persons, settings, and times”(Campbell & Cook, 1979, p. 71). The implications of those limitations for future research results are required to be further considered.
References


119


Silver, M. S. (1987). On the restrictiveness of decision support systems. John E. Anderson Graduate School of Management at UCLA.


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Afrashteh, Sadaf

Title:
The Role of Explanations in Enhancing Algorithmic Fairness Perceptions

Date:
2021

Persistent Link:
http://hdl.handle.net/11343/300172

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.