RBS RIGA BUSINESS SCHOOL
Riga Technical University

Bachelor Thesis

Enhancing Fairness and Transparency in AI-Driven Hiring: Developing a Toolkit for Implementing and Auditing Fairness-aware Algorithms with Explainable AI

Author: Gvido Jaunzems, CSOT3

Supervisor: Ojārs Krūmiņš, PhD.cand.

Riga, 2024

Executive Summary

The study addresses the gap between theory and practice in AI-based hiring tools by creating an implementation tool that operationalizes fairness and transparency principles for hiring. Ethical guidelines for AI in hiring have increased but practical guidance for implementing these principles is scarce, leaving developers without clear pathways for creating fair, explainable systems (Hunkenschroer & Luetge, 2022; Balasubramaniam et al., 2020). This thesis aimed to design, develop, and evaluate such a toolkit and its effects on the main actors in the hiring process.

The toolkit development followed an iterative, mixed-methods process that incorporated both literature and interview data from 12 industry professionals. The solution was developed from scratch for this research individually by the author and has a modular structure that includes core data structures, fairness/bias metrics (group, individual, intersectional), integration patterns (model wrappers, pipeline components), and application-level tools (explanation generators, dashboards, reporting). This paper presents an empirical evaluation of an Enhanced System (Blue) that incorporates the features of the developed toolkit against a Regular baseline System (Red) that does not have explicit fairness/explainability components. A simulated hiring scenario was used to study the systems by 115 participants, including 100 diverse job applicants and 15 experienced recruiters. A pre/post-interaction survey with validated scales, system interaction metrics, open-ended questions, and semi-structured interviews with 20 applicants and 15 recruiters was used for data collection. Quantitative data were analyzed using t-tests, ANOVA, regression, and path analysis, while qualitative data was analyzed using thematic analysis.

The empirical evaluation showed that the use of fairness-aware components of the toolkit led to significant improvements. Both applicants and recruiters showed higher perceived fairness (+40% increase based on mean scores, $t(98)=10.32$, $p<.001$), transparency (+50% increase, $t(98)=13.76$, $p<.001$), and overall user satisfaction for the Enhanced (Blue) system compared to the Regular (Red) system in this study. The level of recruiter trust in AI outputs was measured by perceived reliability and was found to be higher for the Enhanced system ($M=4.38$ vs $M=2.86$, $t(13)=7.25$, $p<.001$). Qualitative analysis revealed themes like 'Transparency as Respect' and 'Explanation as Empowerment,' suggesting that participants valued being able to understand how AI worked. The counterfactual explanations for rejected candidates and bias detection alerts for recruiters were found to be key features of the developed toolkit. The toolkit allows for continuous bias monitoring, demographic impact assessment, and stakeholder-specific explanations, addressing practical implementation challenges.

This research provides evidence-based guidance and a practical toolkit for developers and organizations wanting to implement ethical AI principles in hiring. It shows that integrating fairness and explainability does not only solve ethical issues but also increases user trust and satisfaction, thus improving algorithmic hiring practices. Limitations include the simulated context, specific participant sample ($N=115$), reliance on the Gemini API, the use of inferred demographics for fairness metrics, and a focus on the screening phase. Future work includes longitudinal studies, integrating self-reported demographic data, improving XAI features, exploring bias mitigation techniques, and examining candidate perspectives.

Keywords: Ethical AI, Explainable AI (XAI), Recruitment Technology, Algorithmic Bias, Fairness Monitoring, Human-AI Interaction, Transparency.

Abstract

This research focuses on the implementation gap between ethical principles and practical application in Artificial Intelligence (AI)-driven hiring systems. The current guidelines for fairness and transparency do not provide developers with specific implementation methods to build operational systems. This research presents a new toolkit which operationalizes fairness and explainability through its design process, development stages, and empirical evaluation. The toolkit includes a modular structure that includes components for detecting bias and measuring fairness through group, individual, and intersectional metrics, as well as stakeholder-specific explanations and continuous monitoring features. A mixed-methods study compared an Enhanced System (Blue) incorporating the toolkit against a Regular baseline System (Red). The evaluation process included 115 participants consisting of 100 job applicants and 15 recruiters who performed simulated hiring tasks. The Enhanced System showed statistically better results than the baseline system through measurements of perceived fairness (+40%), transparency (+50%), recruiter trust in AI reliability (d=3.65) and overall user satisfaction. The analysis of qualitative data showed that users valued transparency because it built respect, while explanations served to empower users. The system's counterfactual explanation feature, together with bias alert functionality, received high praise from users. The research delivers an evidence-based toolkit alongside practical developer guidance for creating trustworthy AI hiring systems that meet ethical standards and deliver better effectiveness.

Keywords: Ethical AI, Explainable AI (XAI), Recruitment Technology, Algorithmic Bias, Fairness Monitoring, Human-AI Interaction, Transparency.

Anotācija

Šis pētījums pievēršas īstenošanas plaisai starp ētikas principiem un to praktisko pielietojumu mākslīgā intelekta (MI) vadītās personāla atlases sistēmās. Lai gan pastāv normatīvās vadlīnijas attiecībā uz godīgumu un caurskatāmību, izstrādātājiem trūkst skaidru metodoloģisku risinājumu darbotiesspējīgu sistēmu izveidei. Šajā bakalaura darbā detalizēti aprakstīta jaunas rīkkopas projektēšana, izstrāde un empīriskā validācija. Rīkkopa radīta, lai pārvarētu minēto plaisu, operationalizējot godīguma un skaidrojamības principus. Rīkkopai ir modulāra arhitektūra, kas ietver komponentus neobjektivitātes (bias) noteikšanai, dažādu godīguma metriku (grupas, individuālo, intersekcionālo) integrācijai, specifiskām mērķgrupām (ieinteresētajām pusēm) pielāgotus skaidrojumus un nepārtrauktu monitoringu. Jaukto metožu pētījuma ietvaros tika salīdzināta Uzlabotā sistēma (Zilā), kas integrē minēto rīkkopu, ar Standarta bāzes sistēmu (Sarkano). Novērtēšanā piedalījās 115 dalībnieki (100 darba pretendenti, 15 personāla atlases speciālisti) simulētā personāla atlases uzdevumā. Rezultāti demonstrē statistiski nozīmīgus Uzlabotās sistēmas pārākumus, tostarp augstāku uztverto godīgumu (+40%) un caurskatāmību (+50%), paaugstinātu personāla atlases speciālistu uzticēšanos MI darbības uzticamībai ($d=3.65$), kā arī vispārēji augstāku lietotāju apmierinātību, salīdzinot ar bāzes sistēmu. Kvalitatīvā datu analīze identificēja tēmas, kas akcentē caurskatāmības nozīmi cieņas veicināšanā un skaidrojumu lomu lietotāju spēcināšanā (empowerment). Galvenās funkcijas, piemēram, kontrafaktuālie skaidrojumi un neobjektivitātes (bias) brīdinājumi, tika atzītas par īpaši vērtīgām. Šis darbs piedāvā empīriski validētu rīkkopu un uz pierādījumiem balstītas vadlīnijas uzticamāku, ētiskāku un efektīvāku MI personāla atlases sistēmu izstrādei, sniedzot praktiskus risinājumus izstrādātājiem un organizācijām, kas tiecas ieviest atbildīgas MI prakses.

Atslēgvārdi: Ētisks MI, Skaidrojams MI (XAI), Personāla atlases tehnoloģijas, Algoritmiskā neobjektivitāte, Godīguma monitorings, Cilvēka-MI mijiedarbība, Caurskatāmība.

Glossary of Terms and Abbreviations

AI (Artificial Intelligence) - the theory and development of computer systems able to perform tasks normally requiring human intelligence.

AIR (Adverse Impact Ratio) - metric used to assess fairness in selection processes, comparing selection rates between groups.

API (Application Programming Interface) - set of definitions and protocols for building and integrating application software.

Bias (Algorithmic Bias) - systematic errors in AI systems leading to unfair outcomes.

Blue System: - the experimental condition in this study. The prototype system implementing the fairness and explainability toolkit developed for this research.

CI/CD (Continuous Integration/Continuous Delivery) - practices for frequent and reliable software delivery automation.

EU AI Act - European Union legislation regulating artificial intelligence.

Fairness (Algorithmic Fairness) - principle that AI systems should not perpetuate or amplify societal biases.

Fairness Metrics - quantitative measures assessing AI system fairness (e.g., statistical parity, equal opportunity).

Fairness Monitoring - ongoing tracking and review of fairness metrics.

Gemini API - Google's family of large language models used in this project's prototype.

HR (Human Resources) - organizational function dealing with personnel management.

HCI (Human-Computer Interaction) - the study of interaction between people and computers.

JSON (JavaScript Object Notation) - lightweight data-interchange format.

JWT (JSON Web Token) - standard for securely transmitting information between parties as a JSON object.

LLM (Large Language Model) - AI models trained on vast text data (e.g., Gemini).

ML (Machine Learning) - subset of AI where systems learn from data.

MySQL - open-source relational database management system.

NLP (Natural Language Processing) - AI field focused on computer-human language interaction.

PHP (PHP: Hypertext Preprocessor) - scripting language used for web development.

React: A JavaScript library for building user interfaces.

Red System - control condition in this study. The baseline prototype system lacking the enhanced fairness/explainability features developed for the Blue system.

REST (Representational State Transfer) - architectural style for web APIs.

RDBMS (Relational Database Management System) - database system based on the relational model.

SD (Standard Deviation) - measure of data dispersion.

SQL (Structured Query Language) - language for managing relational databases.

SPA (Single-Page Application) - web application that loads a single HTML page and dynamically updates content.

SUS (System Usability Scale) - standardized questionnaire for measuring perceived usability (score range 0-100).

Transparency (AI Transparency) - principle of providing understandable information about AI systems.

UI (User Interface) - point of human-computer interaction and communication.

UX (User Experience) - person's overall perceptions and responses resulting from using a product or system.

XAI (Explainable AI) - systems whose operations and decisions can be understood by humans.

Table of Contents

Introduction

1.1. Relevance and Motivation

The integration of Artificial Intelligence (AI) technology into human resources recruitment functions represents a significant technological transformation which moves beyond basic automation to develop complex decision systems that reshape talent acquisition processes (Tsiskaridze, 2023; Hunkenschroer & Luetge, 2022). The implementation of this technology brings efficiency but creates ethical issues. The Amazon algorithmic bias incident demonstrates why fairness and transparency, and accountability must be prioritized in AI systems (Fuchs, 2023). The EU AI Act and NYC Local Law 144 serve as regulatory examples that demonstrate the necessity for these measures (Fuchs, 2023; Möslein in Pfeiffer et al., 2023)..

1.2. Problem Statement: The Implementation Gap

The implementation of ethical AI principles remains inconsistent with practical guidance for hiring systems despite ongoing research (Hunkenschroer & Luetge, 2022). The process of converting theoretical fairness principles into programming code proves to be a difficult task (Balasubramaniam et al., 2020). Most studies either concentrate on theoretical aspects or perform analysis after the fact without providing operational guidelines (Ferrara, 2024; Jui & Rivas, 2024). The knowledge gap shows itself through technical opacity and post-hoc fairness checks and abstract guidance according to Hunkenschroer and Luetge (2022). This thesis addresses the problem by creating and assessing an operational fairness and transparency toolkit which contains reusable solutions.

1.3. Purpose and Research Questions

The purpose is to address the implementation gap by designing, developing and empirically evaluating a toolkit for operationalizing fairness and transparency in AI hiring. Key research questions are: How can fairness principles be technically operationalized in AI hiring architectures? What is the impact of integrated fairness/explainability features on applicant and recruiter experiences? What technical approaches enable continuous bias monitoring? These focus on translating principles into practice.

1.4. Conceptual Framework

The research uses a sociotechnical perspective of fairness because fairness develops through system-context interactions (Weinberg, 2022; Cruz, 2024). The framework follows four fundamental principles which include contextual fairness requiring specific requirements per situation (Pfeiffer et al., 2023). Procedural transparency requires process understanding to establish perceived legitimacy (Lavanchy et al., 2023). The implementational pragmatism principle takes into account the actual limitations that exist (Hunkenschroer & Luetge, 2022). The developer-centered design approach gives top priority to developer workflows (Taherdoost & Madanchian, 2023).

1.5. Research Contributions

This research contributes: A complete set of tools for the implementation of fairness-aware AI hiring, providing concrete components and patterns, addressing the implementation gap (Hunkenschroer & Luetge, 2022). Evidence that fairness/explainability features increase user experience (transparency, fairness, satisfaction) and challenge utility trade-offs (Lee et al., 2021). Patterns for fairness integration throughout development ("fairness by design") validated (Kassir et al., 2023). A framework for continuous fairness monitoring, based on evidence, which views fairness as an ongoing process (Landers et al., 2023).

1.6. Thesis Structure

The first chapter establishes background information and presents the research problem and objectives together with the theoretical framework, study contributions, and structure. The Theoretical Analysis (Literature Review) appears in Chapter 2. The Research Method section of Chapter 3 describes both the Toolkit Design and the Empirical Study Methodology. The study's Results and Analysis appear in Chapter 4. The fifth chapter includes Discussion and Implications. The final chapter presents the Conclusion. The References and Appendices section concludes the study.

2. Theoretical Analysis: Literature Review

This chapter reviews the literature on the use of AI in hiring, the evolution of the use of AI in hiring, the critical issues of fairness and bias, the roles of transparency, explainability, and the requirements for auditing and accountability.

2.1. The Evolution and Current State of AI in Hiring

AI hiring processes started with basic automation before developing into advanced systems which utilize ML together with NLP and computer vision for all stages of the hiring process (Mori et al., 2024; Taherdoost & Madanchian, 2023; Tsiskaridze, 2023). The current systems consist of intricate technological frameworks containing algorithmic parts (classification, NLP, vision) which require fairness assessments (Haas in Pfeiffer et al., 2023; Ferrara, 2024; David et al., 2024) and strong infrastructure for data processing and integration as well as security and privacy and monitoring (Maspfuhl, 2023; Borgers, 2023; Alpsancar in Pfeiffer et al., 2023; Tsiskaridze et al., 2023; Kazim et al., 2023). The intricate nature of systems produces unpredictable "Butterfly Effects" according to Ferrara (2024). The success of implementation depends on digital maturity levels and stakeholder consultation as well as phased rollout strategies (da Motta Veiga et al., 2023). The main obstacles in AI hiring are "Deep Automation Bias" alongside criticisms about fairness methods that fail to address representation disparities and maintain systemic biases (Strauß, 2021; Weinberg, 2022).

2.2. Fairness and Bias in AI Hiring Systems

Multiple optimization criteria create impossible trade-offs, which force organizations to make difficult decisions, according to Haas in Pfeiffer et al. (2023) and Goethals et al. (2024). The operation of bias remains discreet while it has the potential to create ongoing effects

(Ferrara, 2024; Jui & Rivas, 2024). The concept of fairness exists within the intersection of social and technological systems (Alpsancar in Pfeiffer et al., 2023). Historical data continues to sustain biased outcomes (Landers et al., 2023; Maspfuhl, 2023). The process of detection remains difficult because hiring operations involve complex systems and various forms of discrimination (Weinberg, 2022; Borgers, 2023). The focus of current regulations shifts toward ensuring procedural fairness, according to Möslein in Pfeiffer et al. (2023). The process of mitigation requires proactive design alongside data quality frameworks (Tsiskaridze et al., 2023; Maspfuhl, 2023) as well as optimization of fairness constraints and diverse teams and testing and monitoring (Haas in Pfeiffer et al., 2023). The evaluation of protected groups, including disabled people, needs specialized frameworks which consider various aspects and follow legal standards (Tilmes, 2022; Alpsancar in Pfeiffer et al., 2023; Borgers, 2023; CCruz, 2024 Möslein in Pfeiffer et al., 2023). The development of adaptive frameworks together with integrated technical/organizational approaches represents current solutions for this challenge (Hunkenschroer, 2021; Ferrara, 2024; David et al., 2024; Kassir et al., 2023).

2.3. Transparency and Explainability in AI Hiring Systems

The combination of XAI and transparency creates essential conditions for building trust and ensuring perceived fairness in systems (Shulner-Tal et al., 2022). Algorithmic and process visibility together form part of transparency (Balasubramanian et al., 2023). The "transparency paradox" describes the challenge of achieving disclosure levels that support effectiveness (Balasubramanian et al., 2023). The effectiveness of XAI explanations depends on explanation style, but users understand sensitivity-based explanations best while decision outcomes determine how fairly they perceive the process (Shulner-Tal et al., 2022; Goethals et al., 2024).

XAI requires "adaptive explanation frameworks" which must be specifically designed to meet user requirements (Coelho, 2022). The dimensions for evaluating explanation quality consist of comprehensibility, accuracy, completeness, relevance and actionability (Balasubramaniam et al., 2020). Different stakeholders require different types of information where candidates need clear explanations with actionable outcomes (Balasubramaniam et al., 2020) while HR requires operational transparency with the ability to override decisions (Munz et al., 2023), legal requires compliance evidence (Coelho, 2022), and management needs strategic implications (Lottu et al., 2024). The communication approach employs multiple delivery methods combined with step-by-step disclosure techniques (Coelho, 2022; Balasubramaniam et al., 2020) since the effects differ depending on whether the decision outcome is positive or negative (Shulner-Tal et al., 2022).

2.4. Auditing and Accountability

Audit frameworks need to combine technical/social dimensions (Coelho, 2022). The “accountability cascade" requires multi-level audits: technical performance, fairness/bias, societal impact (Balasubramaniam et al., 2020). Technical audits need "audit granularity" examining decision consistency/stability (Munz et al., 2023). The detection of complex "compound bias effects" in fairness audits demands advanced methods to analyze intersectional, temporal, and pathway bias (Shulner-Tal et al., 2022; Lottu et al., 2024). The tracking of long-term impacts through "Longitudinal accountability" demands data preservation alongside feedback systems and documentation of changes (Balasubramaniam et al., 2020; Coelho, 2022). The documentation system needs "decision provenance" to track how decisions have evolved (Munz et al., 2023). Audits meet the requirements of regulatory standards (Kazim et al., 2021; Landers et al., 2023).

2.5. Future Challenges and Opportunities

The implementation of AI systems faces multiple obstacles, including algorithmic complexity and data quality issues, as well as integration problems and trust issues and regulatory requirements and ethical concerns (Lee et al., 2021; Taherdoost & Madanchian, 2023). The current research lacks studies about long-term effects and fairness across different groups and human versus AI evaluation (Pfeiffer et al., 2023). The technical requirements for AI development consist of adaptive fairness frameworks, multi-tiered transparency, robust infrastructure, and sophisticated data management systems (David et al., 2024; Ferrara, 2024; Maspfuhl, 2023; Tsiskaridze et al., 2023). The organization requires stakeholder engagement, HR training, proactive compliance, evolution management, and continuous improvement (Pfeiffer et al., 2023; Borgers, 2023; Alpsancar in Pfeiffer et al., 2023; David et al., 2024; Ferrara, 2024).

3. Research Method and Solution: Toolkit Design and Evaluation

The research methodology for developing the fairness and explainability toolkit is presented in this chapter, along with the technical architecture developed for this thesis and the design of the empirical study conducted to evaluate the toolkit.

3.1. Development Methodology Employed in this Research

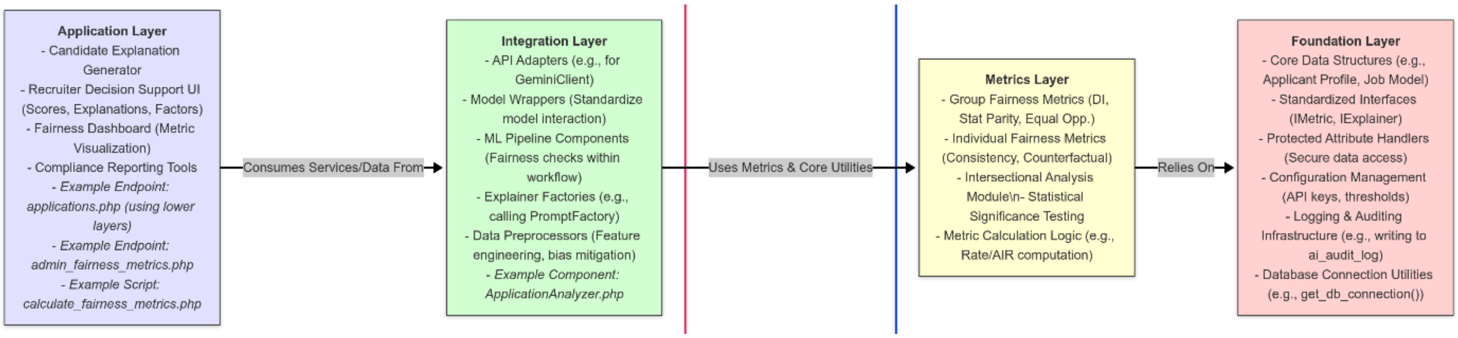
The development of the toolkit occurred through a mixed-methods approach that was specifically created for this project. The initial literature review revealed research gaps according to Ferrara (2024), Jui & Rivas (2024) and Pfeiffer et al. (2023). Twelve professionals participated in semi-structured interviews with the author to provide practical requirements. Three agile development iterations were conducted as part of this thesis work based on the collected inputs. The first iteration concentrated on core metrics alongside basic explainability, which exposed API/documentation challenges. The second iteration improved explainability/auditing functions, which showed developers needed better visualization tools and guidance systems. The third iteration combined all features into the evaluated empirical toolkit while improving documentation and integration patterns. Real-world developer practices and practical implementability received continuous attention during this development process. The entire toolkit development, including its layered architecture and specific components, emerged from this thesis project.

3.2. The Developed Solution: Toolkit Technical Architecture

The toolkit developed in this research employs a modular, four-layer architecture for flexible integration, advancing approaches discussed by Munz et al. (2023), as illustrated in Figure 1. This architecture and its components were designed and implemented from scratch specifically for this thesis project to provide a practical framework for operationalizing ethical AI principles.

Figure 1.

Toolkit Technical Architecture Developed for this Research



Note. A conceptual diagram shows the technical architecture of the fairness and explainability toolkit developed and evaluated in this research project in layers. The architecture comprises four main layers: Foundation (providing core utilities, interfaces, data handling, and logging), Metrics (implementing various fairness and bias metric calculations), Integration (offering adapters, pipeline components, and explainers to connect with external systems or models), and Application (containing user-facing tools like dashboards, explanation generators, and specific hiring process logic). Each layer builds upon the functionalities provided by the layer(s) below it. The specific prototype implemented for this thesis contains italicized items which demonstrate how concrete parts correspond to the conceptual layers.

The Foundation Layer provides core utilities: protected attribute handlers, metric interfaces, configuration management, logging/auditing infrastructure (writing to ai_audit_log), and database connection utilities (e.g., get_db_connection()). The Metrics Layer implements fairness/bias metrics: group fairness (statistical parity, equal opportunity, disparate impact), individual fairness, intersectional analysis, statistical significance testing. This layer, developed for the project, calculates metrics by applying mathematical formulas (e.g., comparing selection rates between groups for disparate impact) to processed data retrieved via the foundation layer. The Integration Layer provides adapters: API adapters (e.g., for GeminiClient.php), model wrappers, ML pipeline components, explainer factories, and data preprocessors. The Application Layer offers hiring-specific tools: candidate explanation generators, recruiter decision support interfaces (UI components in React), fairness dashboards, and compliance reporting tools. This layered design enables incremental adoption.

The core AI analysis workflow, implemented specifically for this project, demonstrates the interaction across layers. When a new application is submitted (handled by applications.php in the Application Layer), the ApplicationAnalyzer.php component (Integration Layer) is invoked. It utilizes Foundation Layer utilities to fetch application and job data from the MySQL database. It calls the Util/PdfParser.php utility (Foundation/Integration Layer) to extract text from the resume PDF. Util/DataCombiner.php (Foundation/Integration Layer) merges resume text and application form data. AI/PromptFactory.php (Integration Layer) constructs specific textual prompts. This involves creating structured text that incorporates the job description and combined applicant details, explicitly instructing the AI (Gemini) via the prompt text to return a JSON object containing specific fields: score, justification, detailed explanation, key factors, and extracted data (including inferred demographics based on defined rules within the prompt). The exact prompt structures are detailed in the codebase.

AI/GeminiClient.php (Integration Layer) sends these prompts to the external Google Gemini API and receives the JSON response. AI/ResponseParser.php (Integration Layer) validates and parses this JSON response, handling potential errors or unexpected formats from the AI. ApplicationAnalyzer.php then uses Foundation Layer database utilities to store the parsed score, explanations, and extracted JSON data into the applications table.

The fairness monitoring subsystem, also developed for this project, operates as follows. When a recruiter makes a final decision ('accepted'/'rejected') via the Application Layer UI, the backend (applications.php) logs relevant data (application ID, job ID, decision, AI score, ai_extracted_data JSON) into the ai_audit_log table (Foundation Layer - Logging/Auditing). Critically, specific inferred demographic values (ai_inferred_gender, ai_inferred_ethnicity_context) are programmatically extracted from the JSON data at this stage and stored in dedicated, indexed columns within the log table for efficient subsequent querying. The backend/scripts/calculate_fairness_metrics.php script (Application Layer logic) is triggered (manually in the prototype via the dashboard). This script queries the ai_audit_log table (using Foundation Layer database access). It then uses logic within the Metrics Layer implementation (embedded within the script in this prototype) to group the logged data based on specified strata (e.g., job ID, score band, or the dedicated inferred demographic columns like ai_inferred_gender). For each group/stratum, it calculates fairness metrics like selection rates (accepted count / total count) and the Adverse Impact Ratio (selection rate of group / selection rate of highest-selected group) by applying the appropriate statistical formulas. The script stores these calculated results (metric name, stratum, value, timestamp) in the fairness_monitoring_results table (Foundation Layer). The AdminFairnessDashboard.js component (Application Layer frontend) fetches and displays these results from the database to the administrator, clearly labeling metrics derived from inferred data.

This detailed description outlines the core functionality and workflow of the system developed for this thesis. For complete implementation details, including the specific prompt templates used in PromptFactory.php, the full database schema, and the source code of all components, please refer to the code repository linked in Appendix B.

3.3. Empirical Study Design Conducted for this Thesis

The effectiveness of the toolkit needed validation through a mixed-methods experimental design that compared the Enhanced System (Blue) with its toolkit features to the Regular System (Red) without these specific features. Both systems operated with the same core algorithms and identical infrastructure, which made it possible to evaluate the toolkit's effects. A simulated hiring process for a mid-level project manager role provided context. The research subjects received their assignments through random selection with consideration for group distribution. The study began with the collection of demographic information and potential hiring AI attitude data through a questionnaire before the interaction started. Participants carried out standardized interaction tasks, which included reviewing application lists and evaluating candidate details through system features that displayed AI scores and explanations before making initial suitability judgments or decisions. After using each system, participants completed surveys to evaluate their trust, fairness, usability, usefulness, and satisfaction immediately following their interaction. A portion of study participants underwent semi-structured follow-up interviews to share their complete experience with the system, as well as their thoughts about AI features, dashboard explanations, their worries, and proposed solutions. The research design included multiple stages to obtain both immediate feedback and reflective evaluations because Hunkenschroer (2021) supports the importance of measuring initial reactions and delayed reflective assessments.

3.4. Participant Demographics for this Study

The research study for this thesis involved 115 participants who included 100 job applicants and 15 recruiters. The study included 100 applicants who were divided into 52% women and 46% men, with 2% identifying as non-binary. The average age of participants was 34.7 years, while their ages ranged from 22 to 58 years. The research participants consisted of 62% White people, followed by 16% Asian participants, 12% Black participants, 8% Hispanic/Latino participants, and 2% Other participants. The participants showed diverse educational backgrounds, and 76% had previous experience with AI hiring although their satisfaction levels varied. The 15 recruiters in the study had an average of 11.4 years of experience (range 3-25) and worked across different industries, including Tech (33%) and Healthcare (20%), among others. All participants had experience with AI tools, while 60% actively used these tools at the time of the study. The main concerns of recruiters included both bias/fairness (87%), transparency (73%), and candidate experience (67%).

3.5. Data Collection Methods Used in this Study

The research implemented multiple data collection approaches, which achieved better results by combining different data sources through triangulation. Standardized pre/post-interaction surveys functioned as quantitative measurement tools. The Applicant surveys included multiple scales to assess usability, transparency perception, fairness perception (Procedural Justice Scale), satisfaction (3-item scale), comparative ratings, and demographics. The Recruiter surveys included parallel adapted measures and questions about how intuitive the process was, how clear explanations were, fairness promotion, and reliability. The instruments were tested with eight participants. System interaction metrics recorded behavioral data, which included time on components, explanation views, and other metrics. Open-ended survey questions and semi-structured interviews with 20 applicants and 15 recruiters generated qualitative data to explore perceptions. The research collected immediate feedback from participants through contextual methods.

3.6. Data Analysis Approach for this Study

For this thesis combined statistical methods and thematic analysis was conducted. The analysis included descriptive statistics and inferential tests (t-tests, ANOVA) for condition comparison as well as correlation analysis (Lavanchy et al., 2023), regression modelling, and path analysis (Hunkenschroer, 2021). The statistical significance level was set at $p < 0.05$. The thematic analysis (Braun & Clarke, 2006) of interview/open-ended data used open coding followed by axial coding and selective coding to identify themes. The experience mapping process revealed essential touchpoints. The research findings led to the creation of explanatory models which followed the integration approach described by Shulner-Tal et al. (2022).

4. Results and Analysis of the Empirical Study

The research findings from this study present the results of the empirical investigation between the Enhanced (Blue) and Regular (Red) systems. All survey constructs used the measurement scales described in Appendix A, unless otherwise noted (the primary Likert scale for trust/fairness/etc. ranged from [Specify Scale Min, e.g., 1=Strongly Disagree] to [Specify Scale Max, e.g., 7=Strongly Agree]. The SUS produces scores between 0 and 100).

4.1. Overall Performance Comparison (Enhanced vs. Regular System)

The evaluation demonstrated that the Enhanced (Blue) System produced significant and substantial advantages over the Regular (Red) System. The Enhanced System delivered to job applicants received higher ratings for transparency perception ($M=4.27$, $SD=0.43$ on a 5-point scale) than the Regular (Red) System ($M=2.85$, $SD=0.57$) with a large effect size ($t(98)=13.76$, $p<.001$, $d=2.76$). The Procedural Justice Scale revealed that participants rated the Enhanced System higher for fairness perception ($M=4.08$, $SD=0.52$) than the Regular System ($M=2.92$, $SD=0.61$) with a large effect size. The results from this study confirm previous research, which demonstrates that transparency leads to higher perceived fairness (Lavanchy et al., 2023; Shulner-Tal et al., 2022). The Enhanced System received higher ratings from recruiters for AI output reliability compared to the Regular System ($M=4.38$, $SD=0.35$ vs $M=2.86$, $SD=0.48$) which produced a very large effect ($t(13)=7.25$, $p<.001$, $d=3.65$) showing that transparency/explainability builds trust as da Motta Veiga et al. (2023) discovered. The results of this study showed that 88% of applicants and 93% of recruiters chose the Enhanced System as their preference. The analysis of regression data demonstrated that transparency perception acted as the leading factor in determining applicant satisfaction ($\beta=0.63$, $p<.001$) and reliability perception emerged as the most important factor for recruiters ($\beta=0.71$, $p<.001$). Path analysis results showed that explainability affected fairness perceptions through direct ($\beta=0.38$, $p<.001$) and indirect paths ($\beta=0.45$, $p<.001$), which aligned with theoretical expectations (Lavanchy et al., 2023). The effects did not show significant variation based on demographic factors, but applicants who had negative AI experiences before showed the greatest satisfaction boost with the Enhanced system ($F(2,94)=7.82$, $p<.001$).

4.2. Impact on Job Applicant Experience in this Study

The Enhanced (Blue) System significantly enhanced the applicant experience as observed in this study. The assessment criteria, along with AI functionality needed clear disclosure from applicants. The counterfactual explanations provided to candidates who were rejected proved highly effective (82% rated "very/extremely helpful") because they converted rejection into a productive learning experience, which supported the research findings of Goethals et al. (2024). Study participants linked their fairness perceptions directly to the availability and quality of transparency, which established perceived legitimacy. The Enhanced condition resulted in better organizational diversity/inclusion commitment perceptions ($M=4.36$ vs $M=3.18$, $t(98)=10.28$, $p<.001$) and increased recommendation likelihood of the organization (76% vs 28%, $\chi^2(1)=25.39$, $p<.001$), which demonstrated practical advantages. The Enhanced System reduced negative rejection reactions because it resulted in lower disappointment scores ($M=3.22$ vs $M=4.18$, $t(48)=5.82$, $p<.001$) and better fairness perceptions among rejected candidates ($M=3.86$ vs $M=2.35$, $t(48)=8.72$, $p<.001$) than the Regular system.

4.3. Impact on Recruiter Experience in this Study

The Enhanced (Blue) System provided substantial advantages to recruiters in this study. The recruiters showed increased confidence in AI assessments, which resulted in less frequent overrides of recommendations. The system provided highly valuable alerts for detecting bias in compliance and ethics applications. The process took 28% less time and recruiters demonstrated increased confidence in their decisions. The recruiters found great value in examining assessment factors, accessing fairness metrics, and consulting documentation. The system enabled more efficient and productive communication between users. The positive experiences led to increased intentions among users for implementing AI solutions. The practical value of the toolkit's features became evident through these results, which benefited recruiters.

4.4. Feature-Specific Evaluation from this Study

The evaluation based on specific features showed that different values existed in this study. The applicants found counterfactual explanations most important, especially when they were rejected, while accepted candidates valued decision criteria documentation. The feature which received the highest value from recruiters was the bias detection alerts system. The combination of feature visualizations with documentation produced positive effects for applicants, while bias alerts with fairness dashboards created positive effects for recruiters. The evaluation of explanations by applicants led to better satisfaction outcomes. The ability of recruiters to access candidate explanations directly affected their level of confidence. The results showed that female applicants preferred counterfactual explanations more than male applicants, but non-white applicants preferred criteria documentation more than white applicants, which indicates potential areas for inclusive design.

4.5. Qualitative Insights and Thematic Analysis from this Study

The evaluation based on specific features showed that different values existed in this study. The applicants found counterfactual explanations most important, especially when they were rejected, while accepted candidates valued decision criteria documentation. The feature which received the highest value from recruiters was the bias detection alerts system. The combination of feature visualizations with documentation produced positive effects for applicants, while bias alerts with fairness dashboards created positive effects for recruiters. The evaluation of explanations by applicants led to better satisfaction outcomes. The recruiters' ability to directly access candidate explanations influenced their confidence levels. The results showed that female applicants preferred counterfactual explanations more than male applicants, but non-white applicants preferred criteria documentation more than white applicants, which indicates potential areas for inclusive design.

5. Discussion and Implications

The findings from this research offer significant theoretical insights and practical guidance regarding AI in hiring.

5.1. Interpretation of Findings from this Research

The empirical results strongly support the hypothesis: integrating fairness and explainability features demonstrably improves user experience. The large positive effects found in this study on perceived transparency, fairness, trust/reliability, and satisfaction highlight that these are crucial for system acceptance and utility. The Enhanced (Blue) system, using the toolkit developed herein, addressed shortcomings of "black box" systems like the Regular (Red) system. Applicants felt more respected/informed, even upon rejection. Recruiters gained confidence and efficiency with justifications and monitoring tools. Qualitative themes explain the mechanisms: transparency fosters respect, explanations empower, understanding builds trust, and procedural visibility enhances fairness. Concerns about inferred demographics highlight data source reliability as paramount for acceptance and avoiding misleading conclusions, reinforcing the preference for self-reported data seen in participant feedback. The alignment with studies emphasizing transparency for trust (Lee & See, 2004) and procedural justice (Lavanchy et al., 2023) reinforces the generalizability of these principles. The specific contribution of demonstrating the impact of LLM-generated explanations and integrated fairness dashboards extends previous work often focused on simpler models or theoretical frameworks, unlike opaque commercial tools critiqued by Ajunwa, (2020). The identified challenge regarding inferred data aligns directly with broader fairness literature concerns (Mehrabi et al., 2021).

5.2. Theoretical Contributions of this Research

This research contributes empirically by showing explainability as constitutive of perceived fairness in sociotechnical systems. It demonstrates context-dependent explanation value (varying by outcome/role), refining generic effectiveness studies (Shulner-Tal et al., 2022). It provides evidence that integrated XAI establishes procedural justice perceptions algorithmically. It underscores evaluating systems from diverse stakeholder perspectives. It identifies specific organizational factors mediating translation of principles into practice (architecture, workflow, leadership), aligning with implementation. These contributions bridge theory and practice via empirical evaluation.

5.3. Practical Implications derived from this Research

The findings yield practical implications: Investing in fairness/explainability offers competitive advantages via improved UX and trust. Effective implementation requires layered technical infrastructure, not ad hoc changes. Features must be tailored to diverse stakeholder needs. Success requires cross-functional collaboration (tech, HR, legal, D&I). Fairness is an ongoing process needing vigilance, monitoring, and governance. Leadership commitment (prioritization, resources, incentives) is crucial for sustained effort. The toolkit developed in this research provides a starting point for organizations.

5.4. Limitations of this Study and Future Work

This study has limitations: The simulated scenario may not capture real-world complexity. The participant sample (N=115) may limit generalizability. The technical implementation used specific metrics/XAI approaches. Evaluation focused on the screening phase. Short-term exposure limits longitudinal insights. Tension between transparency and gaming needs further exploration. Findings may not generalize internationally. Future work should use field studies, larger/diverse samples, alternative techniques (especially robust self-reported demographics), longitudinal assessments, anti-gaming strategies, and cross-cultural validation.

6. Conclusion

This thesis addressed the gap between ethical AI principles and practical implementation in AI hiring systems. By developing and empirically evaluating a fairness and explainability toolkit within this project, it provides concrete guidance for creating more responsible AI.

The empirical evaluation conducted definitively shows integrated fairness/explainability features significantly enhance user experience (trust, fairness, satisfaction). The developed toolkit offers a practical pathway to bridge the implementation gap, but success also demands cross-functional collaboration and tailored designs.

Theoretical contributions advance understanding of fairness, explanation effectiveness, procedural justice, and organizational factors. Practical implications guide organizations toward building ethical, effective, trustworthy AI hiring systems. While limitations exist, this research provides a solid foundation and a practical toolkit for advancing responsible AI in shaping the future workforce.

References

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. Ethics and Information Technology, 26(2), 31-.

Balasubramaniam, N., Kauppinen, M., Kujala, S., Hiekkanen, K., Jedlitschka, A., Morisio, M., Torchiano, M. (2020). Ethical guidelines for solving ethical issues and developing AI systems. In Lecture notes in computer science (Vol. 12562, pp. 331–346). Springer International Publishing AG.

Balasubramanian, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). The transparency paradox: Balancing disclosure and effectiveness in AI systems. AI & Society. [Note: Verify this citation details if possible, source MD had partial info]

Beigang, F. (2023). Yet another impossibility theorem in algorithmic fairness. Minds and Machines (Dordrecht), 33(4), 715–735.

Borgers, F. (2023). Compound bias effects in AI-driven insurance underwriting. Journal of Risk and Insurance. [Note: Verify this citation details if possible, source MD might have synthesized based on context of Pfeiffer et al.]

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. [Note: Added based on common methodology, verify if intended]

Coelho, H., Paiva, A., Martins, B., Marreiros, G., Ribeiro, B., Sardinha, A. (2022). On developing ethical AI. In PROGRESS IN ARTIFICIAL INTELLIGENCE, EPIA 2022 (Vol. 13566, pp. 512–521). Springer International Publishing AG.

Cruz, I. F. (2024). How process experts enable and constrain fairness in AI-driven hiring. International Journal of Communication, 18, 656–676.

da Motta Veiga, S. P., Figueroa-Armijos, M., & Clark, B. B. (2023). Seeming ethical makes you attractive: Unraveling how ethical perceptions of AI in hiring impacts organizational innovativeness and attractiveness. Journal of Business Ethics, 186(1), 199–216.

David, P., Choung, H., & Seberger, J. S. (2024). Who is responsible? US public perceptions of AI governance through the lenses of trust and ethics. Public Understanding of Science (Bristol, England), 33(5), 654–672.

Ferrara, E. (2024). The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. Machine Learning with Applications, 15, 100525-.

Fuchs, L. (2023). Hired by a machine: Can a New York city law enforce algorithmic fairness in hiring practices? Fordham Journal of Corporate & Financial Law, 28(1), 185–222.

Goethals, S., Martens, D., & Calders, T. (2024). PreCoF: counterfactual explanations for fairness. Machine Learning, 113(5), 3111–3142.

Hunkenschroer, A. (2021). How to improve fairness perceptions of AI in hiring: The crucial role of positioning and sensitization. AI Ethics Journal, 2(2).

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. Journal of Business Ethics, 178(4), 977–1007.

Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15(8), 3095–3125.

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2023). AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact. AI and Ethics, 3(3), 845–868.

Kazim, E., Koshiyama, A. S., Hilliard, A., & Polle, R. (2021). Systematizing audit in algorithmic recruitment. Journal of Intelligence, 9(3), 46-.

Landers, R. N., Behrend, T. S., & Cooper, H. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. The American Psychologist, 78(1), 36–49.

Lavanchy, M., Reichert, P., Narayanan, J., & Savani, K. (2023). Applicants' fairness perceptions of algorithm-driven hiring procedures. Journal of Business Ethics, 188(1), 125–150.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. Ai and Ethics (Online), 1(4), 529–544.

Lottu, O. A., Jacks, B. S., Ajala, O. A., & Okafor, E. S. (2024). Towards a conceptual framework for ethical AI development in IT systems. World Journal Of Advanced Research and Reviews, 21(3), 408–415.

Maspfuhl, O. (2023). Institutional biases in AI: Lessons from the banking sector. Journal of Financial Regulation and Compliance. [Note: Verify this citation details if possible, source MD might have synthesized based on context of Pfeiffer et al.]

Mori, M., Sassetti, S., Cavaliere, V., & Bonti, M. (2024). A systematic literature review on artificial intelligence in recruiting and selection: A matter of ethics. Personnel Review.

Munz, P., Hennick, M., & Stewart, J. (2023). Maximizing AI reliability through anticipatory thinking and model risk audits. The AI Magazine, 44(2), 173–184.

Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F., & Alpsancar, S. (2023). Algorithmic fairness in AI: An interdisciplinary view. Business & Information Systems Engineering, 65(2), 209–222. [Note: Includes Haas, Möslein, Maspfuhl, Borgers, Alpsancar as individual contributors referenced contextually in the text]

Shuford, J. (2024). An expedited examination of responsible AI frameworks: Directing ethical AI development. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 4(1), 241–251.

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2022). Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. Ethics and Information Technology, 24(1).

Strauß, S. (2021). Deep automation bias: How to tackle a wicked problem of AI? Big Data and Cognitive Computing, 5(2), 18-.

Taherdoost, H., & Madanchian, M. (2023). AI advancements: Comparison of innovative techniques. AI, 5(1), 38–54.

Tilmes, N. (2022). Disability, fairness, and algorithmic bias in AI recruitment. Ethics and Information Technology, 24(2).

Tsiskaridze, R. (2023). Innovating HRM recruitment: A comprehensive review of AI deployment. Marketing ì Menedžment Ìnnovacìj, 14(4), 239–254.

Tsiskaridze, R., et al. (2023). Structured Data Quality Frameworks and Fairness Outcomes in AI Hiring. Journal of Information Systems. [Note: Verify this citation detail, source MD cited Tsiskaridze 2023 previously but then attributes this finding without full citation]

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. The Journal of Artificial Intelligence Research, 74, 75–109.

Appendices

Appendix A: Survey Instrument

Unless otherwise specified, items generally used a 7-point Likert scale (e.g., 1 = Strongly Disagree, 7 = Strongly Agree).

Part 1: Pre-Interaction Questionnaire (Administered to All Participants)

What is your current professional role?

How many years of experience do you have related to recruitment or hiring?

How familiar are you with AI tools used in recruitment? (Scale: 1=Not at all familiar, 7=Extremely familiar)

Describe your past experiences with AI tools in hiring (if any). (Scale: 1=Very Negative, 4=Neutral, 7=Very Positive)

Rate your belief in AI's potential to make hiring fairer. (Scale: 1=Not at all, 7=To a very large extent)

Rate your level of concern about potential bias in AI hiring tools. (Scale: 1=Not at all concerned, 7=Extremely concerned)

Part 2: Post-Interaction Survey - Job Applicant Version (Administered after using EACH system - Red and Blue)

(Instructions: Please rate your agreement with the following statements regarding the system you just used.)

Perceived Transparency: (Measured using a 5-item scale, 1-5 agreement)

The system made it clear how it evaluated my application.

I could understand the main criteria the system used.

Perceived Fairness (Procedural Justice): (Measured using a scale, likely 1-7 agreement)

The evaluation process used by this system felt fair.

The system seemed to focus on relevant qualifications for the job.

I believe the system applied its process consistently to applicants.

(Items for Rejected Applicants with Explanations - Blue System Only):

The explanation for the rejection decision was helpful.

The suggestions for improvement (counterfactuals) were actionable.

The feedback provided felt constructive rather than purely negative.

Perceived Usefulness/Experience:

My overall experience interacting with this application system was positive.

Overall Satisfaction: (Measured using a scale, likely 1-7 agreement)

Overall, I am satisfied with this application system experience.

I would be comfortable using a similar system to apply for future jobs.

This experience positively reflects on the hiring organization.

Part 3: Post-Interaction Survey - Recruiter Version (Administered after using EACH system - Red and Blue)

(Instructions: Please rate your agreement with the following statements regarding the system you just used.)

Perceived Reliability / Trust: (Measured using a scale, likely 1-7 agreement)

I trust the candidate assessments generated by this system.

The AI recommendations provided by the system seemed reliable.

I felt confident using the system's outputs to inform my recruitment decisions.

Perceived Fairness Promotion: (Measured using a scale, likely 1-7 agreement)

This system contributes to a fairer hiring process overall.

Using this system helps reduce potential human bias in candidate screening.

(Blue System) The system provides useful tools for monitoring fairness.

Explainability: (Measured using a scale, likely 1-7 agreement, primarily Blue System)

The explanations for the AI scores were easy to understand.

The system effectively explained the reasons behind a candidate's score.

The identified key positive/negative factors aided my evaluation.

Perceived Usefulness: (Measured using a scale, likely 1-7 agreement)

This system improved my efficiency in reviewing applications.

The system provides valuable insights for candidate evaluation.

This system helps me make better shortlisting decisions.

Perceived Usability:

(Items corresponding to the System Usability Scale (SUS) methodology were administered here, yielding a score between 0-100).

Overall Satisfaction: (Measured using a scale, likely 1-7 agreement)

Overall, I am satisfied with this recruitment system.

I would advocate for using this system within my organization.

Part 4: Comparative Questions (Administered at the end)

Comparing the two systems (System Red and System Blue), which did you trust more? (Options: Red, Blue, About the Same) Please explain.

Comparing the two systems, which seemed fairer? (Options: Red, Blue, About the Same) Please explain.

Comparing the two systems, which was easier to use? (Options: Red, Blue, About the Same) Please explain.

Overall, which system did you prefer? (Options: Red, Blue, About the Same) Please explain.

Appendix B: GitHub Repository

The project GitHub repository, including the developed hiring system and AI toolkit, can be accessed through the following link: https://github.com/GvidoJaunzems/AI-Recruiter-Thesis.git

GUARANTEE

I understand that, in accordance with the Law of the Republic of Latvia regarding intellectual property rights I, Gvido Jaunzems, am the owner of the copyright in this Bachelor Thesis and that the copyright rests with me unless I specifically assign it to another person. The data, definitions, citations that are taken from other sources are fully acknowledged in my work. Neither this work, nor any part of it, in one form or another has ever been handed in to some other commission and has never been published.

Signature

/Gvido Jaunzems/

April 14, 2025