A COGNITIVE ENGINEERING APPROACH TO TRANSPARENCY OF CONTRASTIVITY OF AI ALGORITHMS

A thesis submitted to Auckland University of Technology in partial fulfillment of the requirements for the degree of Master of Computer and Information Sciences (MCIS)

Supervisor

B.L. William Wong

By

Xeniya Obolonkova

2025

School of Engineering, Computer and Mathematical Sciences

Attestation of Authorship

I hereby declare that this submission is my work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 22 August 2025

Acknowledgements

I would like to sincerely thank my family for their support during my studies. I am also grateful to Mohamed Abouelenien, Associate Professor at the University of Michigan, for providing data for the initial stage of the research; to Aaron Gorrell, President of The Justice Clearinghouse, for granting access to valuable resources on public safety, ethics, and justice technology; to Dmitriy Serebrennikov, Head of the MIND's Department of Sociological Innovations; and to Mikhail Belov, PhD student at the University of California for valuable insights on linguistic anthropology. I also extend my heartfelt thanks to the HCI PhD student group for their support and companionship throughout this work.

I wish to express my sincere gratitude to my supervisor, Professor B. L. William Wong, for his guidance and encouragement. I also extend my thanks to the school administrators for their invaluable support throughout this journey.

Abstract

The increasing adoption of Artificial Intelligence (AI) within critical decision-making domains has intensified the need for transparency, fairness, and explainability in model design and operation. While technical methods for post-hoc explainability have advanced, their integration into system architectures capable of addressing societal, psychological, and governance concerns remains limited. This paper proposes a conceptual framework for AI model transparency that integrates post-hoc interpretability techniques within the principles of Ecological Interface Design (EID)(Vicente, 1995). We validate the applicability of a cognitive engineering approach - specifically, Cognitive Work Analysis (CWA) (Rasmussen, 1985) and Work Domain Analysis (WDA) - to achieve greater model transparency in the area of textual analysis. The framework leverages abstraction hierarchy modelling and constraint visualisation to connect lower-level elements- such as features and coefficients to higher-order functional and relational representations, enabling multi-level reasoning about model behaviour. The approach addresses fairness assessment, bias mitigation, and reasoning quality evaluation for both individual and group predictions, incorporating reasoning in model explanations (Miller, 2018) into "Explanation Contrastivity" metric to make causal reasoning explicit.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation: The need for AI transparency

The current era of automation system evolution is closely associated with the adoption of artificial intelligence (AI). As part of the Fourth Industrial Revolution (Schwab, 2016), AI has achieved "impressive progress, driven by exponential increases in computing power and by the availability of vast amounts of data." While AI systems are transforming industries by enabling greater capacity, faster operations, and more sophisticated decision-making, their non-linear, multilayered architectures often render predictions opaque, potentially undermining credibility. Understanding how a model arrives at its conclusions and assessing the quality of its predictions is critical for various stakeholders: researchers and developers require transparency to improve and troubleshoot models, while end-users demand explainable outcomes to build trust and make informed decisions (Volkov & Averkin, 2024).

Recognizing the centrality of decision-making, scholars emphasize the need to broaden AI governance beyond purely technical considerations, incorporating societal and value-based perspectives and embracing multidisciplinary approaches (Alvarez et al., 2024, p. 1; Ferrara, 2023, p. 2; Narayanan et al., 2024, p. 2). The promotion of transparency, accountability, explainability, fairness, and related ethical principles at the governmental level—supported by academic initiatives—has set a precedent for guiding system evolution toward addressing complex philosophical and societal dimensions. Perspectives such as distributed cognition, joint cognitive systems, and self-organisation view humans and machines as integrated systems,

though each emphasizes different aspects of interaction. Collectively, these perspectives highlight that effective architectures must support not only seamless collaboration between humans and machines but also foster adaptation and resilience in dynamic environments.

Considering human–AI systems within the socio-technical paradigm and developing architectures that enable users to achieve better control over AI systems is an area that has been actively studied over the past decade. As these systems increasingly mirror the complexity of modern environments and rely more on AI in operational processes, it becomes crucial to focus on ways of designing them to provide the transparency and efficiency necessary to build trust between society and technology.

The field of Explainable AI (XAI) addresses this challenge by making algorithms more transparent through detailed insights and justifications for model predictions or decisions (Miller, 2018; U.S. Government Accountability Office, 2021). XAI provides tools for interpreting "black-box" models and addresses ethical dimensions such as explaining predictions (Miller, 2018), striving for equal opportunities (Zhao et al., 2022), interpreting results for reaching fairness (Zeng et al., 2017), analyzing and highlighting crucial factors in the AI decision-making process, and verifying legitimacy through expert-level comparison (Erdoğanyılmaz et al., 2023). These tasks position XAI as a vital area for addressing current challenges in the AI industry.

However, various scientific bodies underscore that the ethical aspects of AI implementation must be addressed not solely from a mathematical or technological perspective, but also through the lens of human cognition and social behavior. This includes integrating intuition (Lopes, 2025), human perception (Nakao et al., 2023, p. 1768; Narayanan et al., 2024), beliefs (Lopes, 2025), psychological inference (Yang et al., 2025), and causal reasoning (Alvarez et al., 2024, p. 31).

The AI industry's evolving challenges—articulated by governmental bodies and reinforced by long-term operational experience—are shifting development requirements away from a purely technical foundation toward more humanitarian orientations. This shift elevates human–AI teaming from a mechanical interaction to a relationship encompassing philosophical, societal, and moral dimensions that reflect human agency rather than machine function.

Determining how best to design visualizations for human–machine teams, particularly in light of the opacity of AI models, adds complexity to system architecture planning. Traditional approaches that rely on predefined user and task characteristics often fall short in complex environments. Research suggests that a constraint-based approach—such as Ecological Interface Design (EID)—combined with selected elements of technique-driven visualization is better suited to managing instability and ambiguity. Such designs should support adaptive, context-sensitive interactions rather than rigid task scripts or fixed roles (Vernon et al., 2002).

1.2 The Cognitive Engineering Approach

In this thesis we propose a cognitive engineering approach to the representational design of AI transparency for showing the contrastivity of black-box AI algorithms, describing how the cognitive engineering method has been operationalised in the development of a visual representation of an AI's explainability model contrastiveness.

Cognitive engineering is an interdisciplinary field of study, established through the foundational work of Don Norman (Norman, 1986) and Rasmussen (Rasmussen, 1983), that is concerned with system design grounded in the principles of cognitive psychology and human factors. It emphasizes the importance of user-centered approaches and promotes deeper inquiry into the nature of human–machine interaction. Cognitive Work Analysis (CWA) is a prominent method, developed within this field by Rasmussen (Rasmussen et al., 1994), that has generally been used for analysis of complex environments such as the nuclear power industry (Rasmussen, 1985), naval operations (Burns et al., 2000), and the control of industrial processes such as milk pasteurisation (Vernon et al., 2002).

CWA creates abstract models of functional relationships between objects of the work domain, control and management functions, and organisational priorities and goals (Vicente, 1995). Functional relationships are identified through the abstraction process, and the semantics of the work domain are then mapped onto a visual representation of the process using a method known as Ecological Interface Design (EID) (Vicente, 1995), which presents important functional relationships of the process, contextualised by information about system invariants, constraints,

and goals.

EID principles address both contemporary AI transformations and human cognitive requirements, demonstrating effectiveness across high-stakes, complex systems. Following EID recommendations, we apply Work Domain Analysis—specifically, the abstraction hierarchy representation—to model all system functions and their interrelations. This method identifies functional properties, including those not directly observable (Vernon et al., 2002). The abstraction hierarchy, first developed by Rasmussen (Rasmussen, 1985) and later expanded for process control and computer maintenance systems (Rasmussen et al., 1994), defines five levels of abstraction (Naikar, 2016, p. 54). We also incorporate the concept of requisite variety, making system boundaries and constraints explicitly visible within the human–machine interface to enhance control.

1.3 Our case study: Explanation Contrastivity

We introduce the "Explanation Contrastivity" metric to support model reasoning through a broader analysis of causality, as discussed in the philosophical and cognitive science literature (Hilton, 1990; Miller, 2018). This metric makes the model's reasoning explicit within the system interface, improving the evaluation of model results. By implementing this metric, we specifically address the goal of model results observability: we reveal hidden functions (via the explanation contrastivity metric) through more abstract representations, where visual elements articulate the underlying causal reasoning.

1.4 How thesis is organised

We conceptually illustrate the architecture of our approach, which incorporates several components from multiple disciplines 1.1. We delineate the process within three primary phases. The first phase incorporates AI-related tasks: employing the dataset, model training, explanation extraction, and fairness assessment. The second phase employs the Cognitive Work Analysis (CWA) and Work Domain Analysis (WDA) frameworks for cognitive engineering to understand AI-engineering strategies for reaching model transparency and explainability. In the third stage, we utilize the Ecological Interface Design (EID) framework, and specifically its semantic mapping

principle, to represent AI system transparency through the user interface.

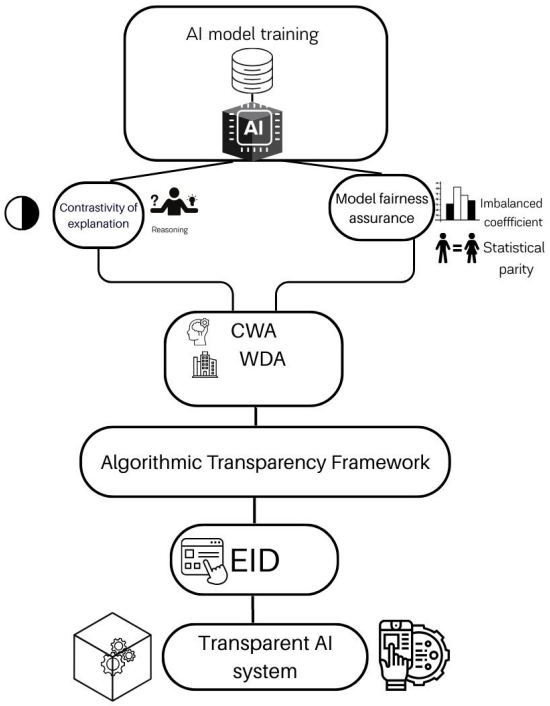

Figure 1.1: Thesis structure diagram

In Chapter II, we review current approaches in AI development from various perspectives, ranging from technical methods to psychological and societal implications. We highlight the emergent need for a multidisciplinary approach in the area of AI governance and ethics to achieve AI transparency.

In Chapter III, we describe the methodology employed in this research, detailing the steps undertaken; the technical and statistical approaches applied to each component of the study; and a brief account of the objectives achieved at each stage within the chosen methods.

Chapter IV relates to the algorithm of the AI system, and demonstrate a domain-specific application the BERT model, which is often described as a "black box" model, trained for occupation prediction based on text features. We provide statistical evidence for the model's prediction and Explanation Contrastivity through a practical AI-model implementation.

In Chapter V, we examine the applicability of Cognitive Engineering approaches to the AI domain. We analyze the activities that characterize the processes performed by specialists during model training and evaluation and define key priorities and goals of the domain. We also analyze the classes of objects encountered in model development and training and propose a multi-level, hierarchical architecture for the system.

In Chapter VI we adopt an AI Transparency framework (Hepenstal et al., 2019) to interpret model explanations against stated goals and constraints. Building on this, we design outcome visualizations using EID principles and propose a multi level, hierarchical interface that maps higher level functions to lower level objects for system transparency. To advance fairness, we introduce the Contrastivity of Explanation metric for assessing explanation quality at both individual and group levels. For effective visualization, we employ Risk Mapper (Wong & Gulden, 2017) to display data or model imbalances and reasoning gaps, supporting transparent evaluation of model behavior.

In Chapter VII we synthesize our findings on AI transparency and reflect on key trade-offs among interpretability, performance, and usability. We note practical benefits and limitations of the proposed approach, outline implications for deployment and governance. Finally, we outline implications for deployment and future research, emphasizing measurement validity, human-AI coordination, and the generalizability of our approach across domains and model families. The research questions we address in this paper are:

i. Applicability of EID principles building transparent AI models?

ii. Applicability of EID principles for model fairness and explainability?

iii. How to utilize principles of cognitive engineering to design an interface concept that supports the process of bias mitigation?

iv. How to evaluate the quality of reasoning in textual model outcomes for fairness assessment?

v. How to Implement the Explanation Contrastivity metric using EID principles to make reasoning quality observable?

1.5 Chapter conclusions

This chapter outlined the thesis aims, objectives, and research questions, and established the theoretical approach and methods used consistently throughout. The next chapter reviews the literature that informs our problem framing and surveys the key theories and technologies underpinning the study.

Chapter 2

Literature review

2.1 Ethical Imperatives for Transparency, Fairness, and Explainability in AI

Ensuring transparency in AI models—particularly in assessing fairness, mitigating bias, and evaluating the quality of reasoning in individual and group predictions—requires careful consideration of ethical implications throughout system development. Enhanced AI systems must proactively address the risk of systematic errors that can result in inequitable or biased outcomes, such as inaccurate predictions, elevated false negative rates, or decisions that disproportionately impact marginalized communities. Such outcomes often stem from biased assumptions embedded during stages like data collection and preparation, highlighting the necessity for ethical awareness in both AI design and deployment.

The use of models that produce unexplainable results can lead to the risk of unexpected consequences. In 2016, a group of investigative reporters and data journalists assessed COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a commercial tool developed by Northpointe, Inc., used for evaluating criminal defendants. The authors compared the likelihood of defendants becoming recidivists based on a questionnaire provided to most individuals booked in jail. They discovered that the system predicts recidivism unevenly between genders, indicating that high-risk women have a much lower risk of recidivism than high-risk men. Additionally, female defendants were more likely to receive a higher score than men when controlling for the same factors. Black defendants had higher recidivism rates overall, and

when adjusted for this difference and other factors, they were 45 percent more likely to receive a higher score than white defendants, including for violent recidivism (Angwin et al., 2016). Incorporating racial and gender biases, the system, as documented by Northpointe Inc., achieved a concordance level of 68%.

During a competition on cancer detection using mammography data, the "Patient ID" feature, which most participants dismissed as irrelevant, unexpectedly proved to have significant predictive power. This occurred because patient IDs indirectly revealed the source of the data, and some sources had imbalanced populations of sick and healthy patients, leading to data leakage that compromised the model's validity. Features being listed during the explanation can be easily observed compared to raw data exploration or analysing the predictions (Ribeiro et al., 2016, p. 2). Undetected data leakage, caused by accidentally mixing training and testing data, can artificially inflate model accuracy while creating a situation where highly correlated features are mistakenly included in the target prediction.

Researchers, governmental bodies, and academic institutions examine AI transparency from diverse perspectives, each shaped by the specific requirements of their domains and offering definitions that align with their respective frameworks and guiding principles. Many studies have shown that issues reported in AI systems can be triggered by a variety of factors (Doshi-Velez & Kim, 2017; Hepenstal, 2023; Miller, 2018; Molnar, 2022; U.S. Government Accountability Office, 2021; Zhao et al., 2022). Effectively mitigating these issues often requires a multidisciplinary approach, drawing on diverse fields and development strategies to build systems that meet the demands for transparency, foster trust, and demonstrate both reliability and efficiency.

Thus, Doshi-Velez and Kim argue that the need for explanation often stems from a lack of proper problem formalization, rather than originating solely from the technical aspects of AI algorithms; this may influence both optimization and system evaluation (Doshi-Velez & Kim, 2017). Additionally, the authors differentiate between qualified variance issues, arising from data, and unqualified bias, linked to the model selection process.

Christoph Molnar, in *Interpretable Machine Learning*, explores the social aspects of the need for explanations. He emphasizes that the necessity arises not only from the desire to know

what is predicted but also why the prediction was made. This is because making a prediction only partially solves real-world problems. Understanding the reasoning behind predictions helps in finding meaning in the world and complementing a mental model of the environment, especially when something unexpected happens (Molnar, 2022). When the model's prediction has no significant impact, the demand for interpretability decreases. For instance, predicting an individual's vacation plans based on social network data may cause little to no trouble, making the interpretation of the result redundant. In contrast, in business or critical industries—where the failure to ensure the accuracy of predictions can lead to substantial losses or pose risks to human life and health—interpretability becomes essential. Molnar urges distinguishing among real-world problems to justify interpretability needs, emphasizing consequences of unexpected model results and the completeness of mental models across application contexts (Molnar, 2022).

Studying AI transparency from the governmental perspective provides a broader view of critical aspects such as transparent data usage, which helps prevent rights violations (U.S. Government Accountability Office, 2021, p. 6). It also highlights the importance of establishing rules for the use of high-quality data to mitigate bias, as well as requirements for ongoing monitoring during system development and deployment. Furthermore, proactive compliance processes are essential for creating regulations and standards that help prevent discrimination and protect privacy (U.S. Government Accountability Office, 2021, p. 30).

2.2 Cognitive Engineering for System Safety

The Cognitive Engineering approach to system design and evaluation emphasizes user-centeredness, the analysis of human factors, the psychological aspects of human–machine interaction, and the identification of crucial relationships within the work domain. It treats performance in complex, safety-critical settings as the joint product of technical, human, organizational, and environmental layers, grounding analysis and design in stable constraints and the semantics of the work domain rather than device features alone (Vicente & Rasmussen, 1990). Within this perspective, Cognitive Work Analysis (CWA) provides a structured, constraint-focused method for making work demands explicit so that information requirements and interface representations can be derived in a technology-independent way. Central to CWA is the Abstraction Hierarchy—a

five-level, means–ends model spanning functional purpose, abstract function, generalized function, physical function, and physical form—which supports both top-down goal decomposition and bottom-up symptom tracing for diagnosis, planning, and control under uncertainty (Rasmussen, 1985). Applicability is demonstrated in shipboard command-and-control, where a Work Domain Model captured mission goals, environmental and tactical constraints, and functional structure, then drove display requirements that coordinate attention and decision making across a team in time-pressured, distributed operations (Burns et al., 2000). A complementary operationalization appears in *Ecological Interface Design for Pasteurizer II*, which documents step-by-step semantic mapping: starting from work-domain constraints (e.g., safety limits, conservation relations) and relational invariants, selecting configural forms that preserve those relations perceptually, and assigning graphical encodings (geometry, orientation, proportion) so operators can directly perceive functional state and emerging anomalies (Vernon et al., 2002, pp. 223–225). Taken together, these cases show how abstract constraints and means–ends links become the geometry of the display, how configural representations elicit emergent features for rapid anomaly detection and diagnosis, and how a shared hierarchical model supports coordination and adaptive control in correspondence-driven domains (Rasmussen, 1985; Vicente & Rasmussen, 1990).

Centering safety, Vicente (1999) argues that a constraint-based, sociotechnical analysis is foundational: designers must begin from environmental and organizational constraints—regulatory limits, physical laws, coupling to markets and teams—so that interfaces, roles, and procedures keep action within safe operating envelopes (Vicente, 1999, p. 12). The Ontario Hydro case (seven reactors shut down primarily for management shortcomings) underscores that safety failures often emerge from organizational control, not component malfunctions (Vicente, 1999). Practically, safety is strengthened when work-domain invariants and boundaries are made perceptually available—through hierarchical modeling that links purposes to physical form (Abstraction Hierarchy) and through displays that reveal functional structure and emergent features for anomaly detection (Bennett & Flach, 1992; Rasmussen, 1985; Vicente & Rasmussen, 1990). CWA's emphasis on constraints also supports safe adaptation across skill-, rule-, and knowledge-based control, clarifying where slips, rule-based mistakes, and knowledge-based errors arise (Rasmussen, 1983). Operationalizing this stance via semantic mapping ensures safety-critical limits and resources are visible at the point of control (Vernon et al., 2002),

aligning with broader evidence that organizational and system-level controls—not just human or technical fixes—are decisive in preventing organizational accidents (Vicente, 1999).

Cumulatively, the cognitive engineering perspective—via CWA's constraint-led analysis, the Abstraction Hierarchy's means–ends structure, and EID's semantic mapping—provides the safety backbone for system design by making limits, invariants, and couplings perceptible at the point of control and supporting adaptive behavior across SRK modes (Bennett & Flach, 1992; Rasmussen, 1983, 1985; Vernon et al., 2002; Vicente & Rasmussen, 1990). Building on this foundation, the next sections turn to complementary approaches for safe and ethical AI, including dataset governance, bias and fairness assessment, explainability, accountability and oversight, and compliance processes that institutionalize transparency and equity alongside safety.

2.3 Technical Approaches to Model Explainability

The technical approaches to achieving model explainability and providing insights into model reasoning can be distinguished by various categories. These include surrogate methods, gradient-based methods, attention mechanisms, and MLP-based explanations (Volkov & Averkin, 2024). Some authors further differentiate methods by their applicability to different kinds of models, distinguishing between model-agnostic and model-specific techniques, as well as by the level at which these methods operate—such as during input processing or prediction (Zini & Awad, 2023). In addition, some authors separate post-hoc explanations derived after a prediction is made from those that produce inherent interpretability.

The authors of the enhanced, frequently applicable post-hoc model-agnostic approach technique for explaining predictions of machine learning models, Local Interpretable Model-agnostic Explanations (LIME), describe the "explanation of prediction" as a process of deriving a "qualitative understanding of the relationship between instance components, such as text, patches, and images." (Ribeiro et al., 2016, p. 1). Applicable explainability techniques are presented in various studies and aim to improve model understandability for non-technical staff (Hepenstal, 2023, p. 3), enhance transparency to support expert reasoning (Hepenstal, 2023,

p. 1), and expose the reasons behind AI-made decisions to individuals to avoid violation of data privacy, discrimination, and bias (Erdoğanyılmaz et al., 2023, p. 1). Reaching LLM explainability requires extensive expertise in AI engineering, involving the use of surrogate models or the development of structure-based explanations that incorporate self-attention mechanisms and other intermediate representations (Zeng et al., 2017).

Castelnovo et al. used LLM-generated explanations to enhance banking managers' interactions with clients by explaining AI-generated recommendations (Castelnovo et al., 2024, p. 10). The authors compared the traditional rule-based approach with LLM-based explainability and evaluated semantic and syntactic similarities between human-validated benchmark text and AI-generated text using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) technique.

The application of SHAP, a local-agnostic method, was demonstrated to extract explanations from various models, such as BERT, RoBERTA, and DistilBERT, for generating explanations of court decisions (Erdoğanyılmaz et al., 2023). The author identified groups of words that influenced predictions based on a certain threshold level and used an Explainability Score based on BLEU and ROUGE metrics to evaluate the similarity of the model-agnostic approach to results highlighted as important by a legal expert. More broadly, SHAP formalizes additive feature attributions with solid theoretical guarantees (Lundberg & Lee, 2017).

A significant shift in developing the interpretability of model predictions, made by Zeng et al., focused on exploring interpretability across a wide range of recidivism prediction problems and comparing various models, including SLIM, Lasso, Ridge Regression, and CART (Zeng et al., 2017). The authors address interpretability through hierarchical logical interpretation for tree-based models and coefficient-based explainability for models with a linear nature. In their work, the authors differentiate between types of models and the associated approaches for representing how each model type links specific features to the model outcome. For the CART model, hierarchical interpretation is applied; in this approach, explanations and reasoning depend on prior levels in the hierarchy, which often makes the relationship between variables and predictions less direct or transparent. In contrast, for linear models, the relationship between input variables and predicted outcomes is represented by the sign and magnitude

of the coefficients, which makes interpretation more straightforward. However, the authors note that both methods of achieving model interpretability can be applied to different types of domains and tasks, depending on whether a logical and structural explanation is preferable to coefficient-based explainability metrics.

2.4 Multidisciplinary Perspectives on Ethical AI

Recent years have brought a notable shift in the requirements for AI adoption, reflecting the need for changes in how such technologies are developed, integrated, and governed. While the framing of the Fourth Industrial Revolution (4IR) has been contested, Moll argues that there is insufficient evidence of extensive changes in work processes, as well as in the social, labor, and economic sectors, to classify the present era as a distinct 4IR (Moll, 2022, pp. 7–8). Referencing early AI and machine learning literature from 1950 to 1997, Moll contends that many of the underlying technological foundations belong to the earlier Third Industrial Revolution (Moll, 2022, pp. 7–8). Nonetheless, over the past decade, AI technologies have been deployed with increasing frequency, giving rise to new challenges, questions, and practices across governmental, academic, economic, and legal spheres.

These developments have accelerated the emergence of a multidisciplinary approach to AI governance. The complexity of contemporary AI systems and the societal impact of their decisions demand that governance frameworks extend beyond purely technical and mathematical considerations. Scholars highlight the necessity of incorporating societal values, philosophical reasoning, and ethical principles alongside engineering and computational expertise (Alvarez et al., 2024; Ferrara, 2023; Narayanan et al., 2024). Fairness, transparency, accountability, and related ethical imperatives—promoted through governmental policies and supported by academic research—are shaping the trajectory of AI system design toward integrating philosophical and societal dimensions with technical performance.

Thus, Yang et al. formalized the concept of inference within the context of human–machine teaming, positioning it as the central component of a Machine Theory of Mind; this framework aims to establish a shared representation between humans and machines to enable more effective

collaboration (Yang et al., 2025). Drawing on a theoretical framework concerning fairness perceptions, Narayanan et al. synthesize empirical findings on perceived fairness in AI and foreground the subjective dimensions of fairness that warrant attention in AI deployment (Narayanan et al., 2024). Lopes, in his exploration of biases in AI systems, emphasizes the importance of individual perceptions and beliefs, noting that people may perceive identical outcomes differently depending on the type of decision-making procedure employed—whether adversarial or inquisitorial—and highlighting the current unpreparedness of AI systems for widespread implementation in the judicial domain due to potential negative impacts on perceptions of procedural fairness (Lopes, 2025). As Lopes underscores, individuals are more likely to accept and comply with decisions when they believe the decision-making process itself to be fair (Lopes, 2025).

A study on the generalization of user-interface requirements—embodied in FairHL, a UI-centered HCAI design for examining model fairness—revealed fundamental differences in the requirements for ML model fairness assessment from the perspective of operators in different roles. These differences arise from variations in how model fairness is perceived. According to the authors, subjectivity is a factor considered by operators in both roles when assessing the fairness of model outcomes. For loan officers, this subjectivity stems from potential data inconsistencies, prompting them to focus on the assessment of individual applications. In contrast, data scientists emphasize group fairness, recognizing that subjectivity can emerge at the level of individual application evaluation and asserting that it is impossible to precisely determine the degree of similarity between individual cases. For both groups, the authors highlight the importance of providing explanations and contextual support to facilitate diagram interpretation by domain experts and scientists engaged in AI fairness assessment (Nakao et al., 2023).

Ferrara (2023) reviews the ethical dimensions of AI deployment, including individual fairness, transparency, and accountability, and synthesizes strategies for mitigating bias; given the challenges of bias mitigation—particularly in generative AI systems—he advises prioritizing more transparent algorithms in which bias can be more easily identified and monitored (Ferrara, 2023).

A broader analysis of AI outcomes aimed at achieving trustworthiness and building reliable AI systems focuses on making these outcomes understandable and providing clear reasoning for decisions made by the model. Explainability involves making algorithms more transparent by offering detailed insights and justifications for a model's predictions or decisions (Miller, 2018; U.S. Government Accountability Office, 2021). Tim Miller defines the explainability and interpretability of autonomous systems as the "ability to generate decisions with ascertaining how well a human could understand the decisions in a given context" and the "ability to explain decisions to people" (Miller, 2018, p. 3). He emphasizes that explanations are contextual, not merely representations of associations and causes (Miller, 2018, p. 7), and argues that extracting statistical relationships and referring to probabilities do not always produce the most effective explanations (Miller, 2018, p. 5). Additionally, his work highlights that explanations are applicable across various AI subfields and identifies explainable AI as a "human–agent interaction problem" (Miller, 2018, p. 5).

This broader framing creates the foundation for examining AI implementation through expanded perspectives, including the social and psychological dimensions essential to achieving transparency, explainability, and effective bias mitigation.

2.5 Cognitive Engineering Methods and Application to AI Transparency

In studying the processes and approaches to building AI systems from a higher-level perspective and considering the system as a socio-technical artifact, authors mention the challenge of designing visualizations for human–machine teams in complex environments (Tieu & Naikar, 2022). The prevalence of rapidly evolving situations, instability, and unpredictability poses difficulties and reveals the limitations of current approaches to information visualization. The author divides approaches into technique-driven visualization, problem-driven visualization, user-centered design, and EID, and notes that work domain analysis serves as the basis for EID. However, the author highlights that designs built on EID principles incorporate a constraint-based approach, which better suits the design of visualizations to accommodate the challenges of complex system environments.

The problem of ill-defined tasks, the problematic and techno-centric way of viewing system design, combined with increasing system complexity and the continual growth in AI utilization, exposes a range of risks for the potential use of such systems. Naikar et al. incorporated three contemporary perspectives from sociotechnical systems—distributed cognition, joint cognitive systems, and self-organisation—integrating these with cognitive work analysis, and demonstrate how these approaches and frameworks can shape the design of human–AI systems (Naikar et al., 2023). It is stated that in rapidly evolving situations, where there is significant uncertainty about present conditions and it is not easy to navigate, the principle of self-organisation, which increases the capacity to deal with uncertainty, is regarded as an effective way to stay in control. This signifies the need for human–AI systems to be self-organising and able to handle situations where the settings are not pre-established. Contextualisation and collaboration are mentioned as important aspects to address in human–AI system teaming; rather than concentrating on human–machine dyads and single, isolated tasks, the goal is to maximize the system's adaptive capacity or resilience. The author suggests identifying system constraints and affordances to determine fields of possible action (Naikar et al., 2023, p. 1689), aiming for an enhanced design capable of operating successfully in dynamic and uncertain situations.

An explicit example of achieving system transparency through CWA application is demonstrated by Hepenstal et al. (Hepenstal et al., 2019). In their work, the authors employ both CWA and CTA approaches to examine specific aspects of the cognitive processes of experts in intelligence analysis, integrating these insights to enhance the transparency of a conversational agent. Notably, this study exemplifies the explicit treatment of the system as a socio-technical instance, where both human and technical factors are considered together. The authors emphasize the importance of identifying system goals, invariants, and constraints as crucial aspects for designing complex systems and ensuring effective human–machine collaboration. Their research addresses dimensions such as transparency in information retrieval, clarity of reasoning, and support for user sense-making within the system.

While the reviewed papers demonstrate significant advancements in technical approaches and system-level strategies for AI transparency, we observe a notable gap. Specifically, there is a lack of studies examining the applicability of EID principles in the development of AI systems

to ensure model transparency and explicitly address model fairness. Most existing research has focused on improving the interpretability of AI through technical explainability methods, post-hoc analysis, or by supporting transparency at the level of information retrieval and system–user interaction. However, these works tend to concentrate on surface-level transparency rather than systematically integrating human-centered design principles, such as EID, into the core processes of model development. As a result, questions remain about how EID might contribute not only to transparency but also to the critical issue of fairness within AI models. This gap highlights the need for further research that brings EID principles into the heart of AI design, aiming to achieve greater trust, accountability, and equitable outcomes in AI systems.

Chapter 3

Methodology

This study investigated the use of cognitive engineering approach for designing the transparency of AI models in the domain of textual recognition. Our investigation involves three broad steps.

3.1 Building the algorithm

At this phase of the research, we train BERT—a model type often referred to as a black-box due to its opaque nature and hidden layers. We employ this model for textual recognition and perform multi-label classification. The dataset used in this study contains short textual samples—job and job-related daily routine descriptions—paired with occupation labels. Because of the opaque nature of the AI model employed for textual classification, we then explore the BERT model's outcome explainability through a post hoc explainer based on Shapley values (Lundberg & Lee, 2017). We extract the textual features associated with specific model outcomes and analyze the quality of the explanations. To measure explanation quality, we employ an approach based on the Explanation Contrastivity metric (Miller, 2018, p. 16) and demonstrate how contrastivity can be used as an indicator of fairness of a system used to analyse male/female discrepancies in various vocations. For analyzing the biased model outcomes, we employ Statistical Parity (SP) (Barocas et al., 2023) and the Imbalance Coefficient (Haibo He et al., 2008, p. 1322), approaches frequently utilized for assessing model outcomes.

We extract Explanation Contrastivity metrics for each prediction and evaluate how these metrics reflect the content and balance of the data, thereby shedding light on the internal

processes involved in the model's decision-making. The expectation is that contrastivity metrics can complement traditional model explanations. Unlike feature-importance explanations—which often require a high level of domain expertise (Zhao et al., 2022)—the contrastivity metric offers a more intuitive and accessible interpretation. This not only demonstrates which features are important but can also indicate potential issues in the model's internal mechanisms or highlight problems related to data preprocessing or imbalances in the dataset itself.

3.2 Analysing, identifying and modeling of key functional relationships

At this step we explore a potential approach of reaching the model transparency by making events of model unfairness and reasoning inconsistency visible. For this purpose, we employ CWA (Vicente, 1999), which help identify the key functions, priorities, and values of the system, and structure the system architecture so that lower-level objects can be observed through higher-level system goals to achieve the main goal of model transparency.

Within the WDA, we conducted an Abstraction-Decomposition Analysis that combines a means-ends analysis with a part-whole (decomposition) analysis. This process yields an Abstraction Hierarchy-spanning functional purpose, abstract function, generalized function, physical function, and physical form-and, critically, surfaces the key functional relationships and dependencies that govern system behavior (Rasmussen, 1985; Vicente, 1999). For rigor and traceability, it is important to explicitly report and describe the resulting hierarchy and the identified relationships (e.g., conserved quantities, constraints, performance measures), as these artifacts provide the analytic rationale for subsequent design decisions and enable lower-level data objects and mechanisms to be interpreted in terms of higher-level goals such as fairness, explainability, and transparency.

3.3 Designing the visual representation for algorithmic transparency

To perform further system architecture planning, we incorporate an Algorithmic Transparency Framework that allows us to interpret the model's outcome explanations through verification of goals and constraints (Hepenstal et al., 2019).

In the next phase of the research, we develop a visualization of the AI model's outcomes by combining various metrics through EID principles (Vernon et al., 2002). Additionally, we analyze the different objects that specialist operates during AI model development and training and form a multi-level, hierarchical architecture for the system interface (Rasmussen, 1985; Vicente, 1999). This interface aims to make the system transparent by allowing the observation of lower-level objects through higher-level abstract functions, mapping them onto the geometry of the user interface. To achieve better model fairness, we present the "Contrastivity of Explanation" metric, which assists in evaluating the quality of explanations for individual predictions as well as for groups (Miller, 2018, p. 16). This signals model unfairness and provides insights into the reasoning behind model outcomes. For effective visualization, we utilize Risk Mapper, a scatter-plotting technique presented by Wong & Gulden for financial risk analysis, to display dataset and AI model training imbalances and reasoning gaps in service of model transparency as the final system goal (Wong & Gulden, 2017).

To our knowledge, we are not aware of other research that applies approaches for achieving model transparency through investigation of reasoning combined with fairness of the model's outcomes, incorporated with a cognitive engineering approach. This makes the topic valuable for building explicit systems within the industrial and societal requirements of AI model transparency, fairness, and reasonable explanation.

Chapter 4

Building of the algorithm for textual classification of occupation

This chapter describes a textual classification task performed using BERT, a Transformer model that excels in a wide range of contextually sensitive language tasks. Its strength lies in its ability to detect subtle changes in meaning based on context, allowing it to identify nuanced differences in phrasing. The advantages of this approach include the model's strong performance even with limited training data (Vaswani et al., 2023) and its reduced need for extensive data preprocessing (Barsever et al., 2020, p. 2). For the task of text classification of current study author utilized the publicly available BERT model from the Hugging Face library: https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#transformers.BertModel.

In the current BERT implementation, the pretrained model is complemented with an additional classification layer for fine-tuning of all parameters on a subsequent task. For the fine-tuning, the BERT model is initially loaded with pre-trained parameters, and then all parameters are adjusted using labelled data, classified as a specific occupation of each class, associated with the downstream tasks. We specifically configured the model hyperparameters for our task, as it has been previously reported that small datasets are sensitive to hyperparameter selection (Devlin et al., 2019, p. 3).

We use a dataset of 10,000 textual samples—job and job-related daily routine descriptions—paired with occupation labels to teach a BERT-based classifier, treated as a black box, to recognize

language patterns linked to jobs. Throughout the chapter, we analyse the explanation of the model's outcome and perform the analysis of Explanation Contrastivity (Miller, 2018). We show how explanations and contrastive analysis help us understand the model's reasoning and assess fairness, turning raw text and predictions into transparent, interpretable insights that guide responsible deployment.

4.1 Data preprocessing and model training

To train a model for classifying text according to occupation labels, we sampled 10,000 records from the original train dataset while maintaining a similar balance of occupations and genders. For testing purposes, we employed 10,000 records from the test dataset, as extensive testing allows us to analyze the explanation contrastivity thoroughly and obtain findings supported by strong quantitative evidence.

This approach enables us to assess contrastivity metrics more efficiently, avoiding the time-consuming process of training a pre-trained BERT model on a full dataset. By focusing on a smaller training subset, we can more directly identify features important for predicting specific occupations and extract the patterns from the model's explanations that are meaningful, while staying focused on the exploration of explanation contrastivity in relation to the model's internal data processing mechanisms. Although the data utilized for training were manually sampled, we preserved the original gender imbalances present in specific occupations (De-Arteaga et al., 2019, p. 3), thereby retaining any inherent bias from the initial dataset (see Figure 4.1).

The model was trained for 3 epochs with a batch size of 8 and a learning rate of 2×10^{-5}. The achieved accuracy was 81%. The classification report for the model's prediction performance is presented in Table A.1 in Appendix A). We implement all models using the `PyTorch` deep learning framework, leveraging the `Hugging Face` *`transformers`* library for pretrained architectures, tokenization, and training utilities. For our task, inputs are tokenized with `AutoTokenizer` and processed by a pretrained transformer. We fine-tune the model end-to-end using the `AdamW` optimizer and evaluate using task-appropriate metrics. This setup enables reproducible, efficient transfer learning while maintaining compatibility with PyTorch's

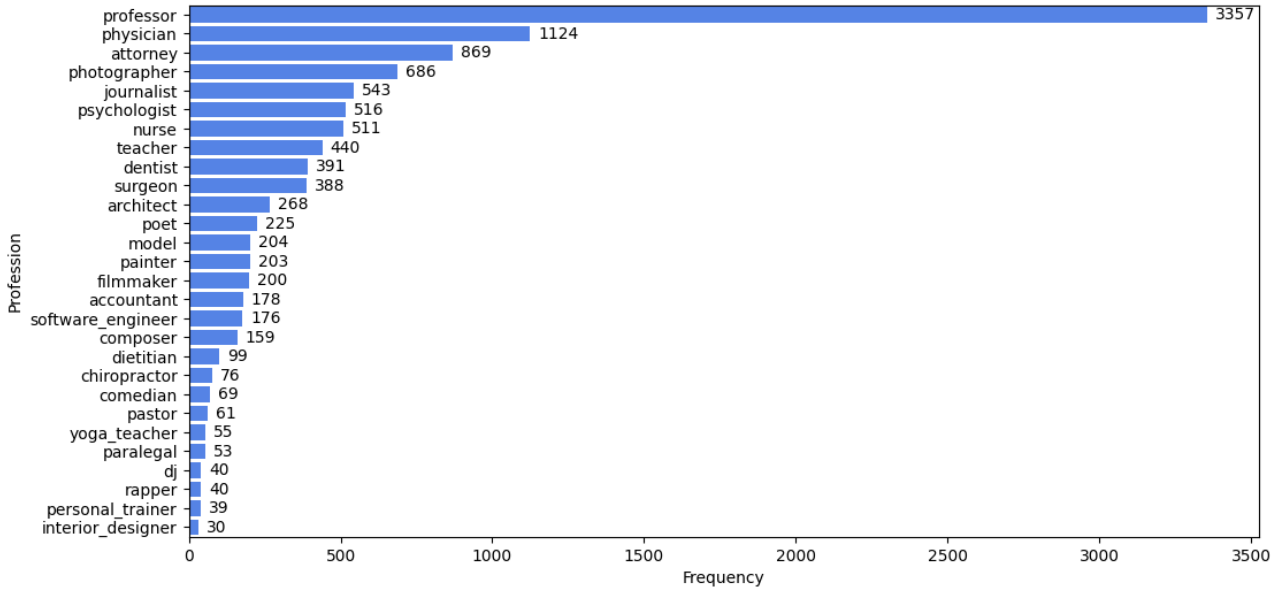

Figure 4.1: Frequency of professions in the set of data

training loops and hardware acceleration. Complete source code and configuration files are available at `github.com/XeniyaVl-git/AI_Transparency/tree/main/ipynb` (release `v1.0.1`), archived at DOI: `10.5281/zenodo.16925254`. .

4.2 BERT Prediction Explanations with Contrastivity Analysis

SHAP offers a compelling approach for explaining model predictions, grounded in solid theoretical foundations (Lundberg & Lee, 2017). It is particularly attractive for analyzing variable importance because it reveals how features relate to the model output through consistent, additive attributions (Lundberg & Lee, 2017).

This phase of the research focuses on applying SHAP to interpret BERT within a transparency framework grounded in cognitive engineering. Prior work supports SHAP's effectiveness for explaining text-model predictions (Erdoğanyılmaz et al., 2023; Volkov & Averkin, 2024; Zini & Awad, 2023). Moreover, SHAP's formulation enables probing and quantifying model reasoning in a principled way, contributing to a more transparent AI system (Lundberg & Lee, 2017).

While complex neural models often achieve the best performance, their intricate nature makes them difficult to interpret, even for experts. SHAP addresses this challenge by leveraging

game-theoretic principles to identify the influence of specific features on each individual prediction (Lundberg & Lee, 2017, p. 3). This creates a trade-off between achieving high accuracy and maintaining model understandability; the ability to correctly interpret a prediction model's output is crucial for fostering appropriate user trust, guiding model improvement, and supporting understanding of the underlying process (Molnar, 2022).

4.2.1 Shap values extraction

Using `shap.Explainer` from the SHAP Python library, we decompose each prediction into additive positive and negative feature contributions; SHAP values indicate how much each feature pushes the model's output toward the positive or negative class for a given instance (Lundberg & Lee, 2017). In the visualization, each Shapley value can be interpreted as a "force" that increases (positive value) or decreases (negative value) the prediction, with the net balance aligning at the instance's actual output (Lundberg & Lee, 2017; Molnar, 2022). Table A.2 in Appendix A presents an example of the top-10 SHAP values extracted for a single instance.

4.2.2 Exploring the Shap Values contrastivity

Our explainability experiments focus on analyzing and quantifying the contrastivity of explanations — an approach originally developed by (Hilton, 1990) and later explored by (Miller, 2018, p. 16), who confirmed its relevance and applicability in the field of XAI. We developed a metric that numerically illustrates the balance between the set of features important for the predicted label (the "present case" in the context of contrastive explanation) and the set of features relevant for the opposite label, which was not predicted (the "contrast case" in Hilton's approach).

To measure the contrastivity of explanation we employed a ratio-based approach — a variation of the well-known Statistical Parity (SP) measure, which was previously used by (Zhao et al., 2022) to assess model fairness.

In this set of experiments, we examine the contrastivity of explanations across different groups of predictions to evaluate how this metric fluctuates for specific prediction groups. We also explore its association with issues in model training and evaluation, such as class imbalance and model bias.

We begin analyzing and measuring the explanation's contrastivity (Miller, 2018, p. 16) by calculating the ratio between positive and negative SHAP values (Lundberg & Lee, 2017). This ratio is used to define the degree of contrast in the explanations. As Miller stated, the cause of an event is usually explained in relation to another event that is counterfactual to the first one (Miller, 2018, p. 16). Counterfactuals are hypothetical outcomes that offer alternative scenarios to what actually occurred, and are referred to by the author as a core concept of the contrastive explanation phenomenon. This relates directly to the case of text classification, as SHAP values are calculated for already predicted instances and offer counterfactual explanations by presenting possible scenarios that could have led to different predictions (Lundberg & Lee, 2017).

Similar approach of identifying biases in the model's outcomes was utilized by Zhao et al. (2022). The authors measured differences in the model's explanations and defined this as an explainability fairness metric. The core concept of the ratio-based explanation discussed in the paper involves determining the percentage of high-quality explanations by focusing on features with the greatest impact on the model's predictions and comparing it using the well-established Statistical Parity (SP) technique(Zhao et al., 2022, p.3). This metric is applicable to assess fairness either by examining prediction fairness across subgroups within a single model or by comparing fairness between two models. The authors extended this method to evaluate explanation quality, defining positive explanations as high-quality explanations, and assessing whether the quality of explanations varies between two sensitive groups. This approach addresses fairness by calculating the proportion of high-quality explanations for subgroups, where a smaller proportion indicates a fairer model.

While ensuring fairness based on high-quality positive explanations has proven effective, it is essential that models are calibrated for fairness in both predictions and explanations, derived from a model-agnostic approach. However, fairness measured solely on positive predictions is not always applicable for certain models used in critical industries. For instance, in (Zeng et al., 2017) study on recidivism prediction models, fairness requirements can vary depending on the application. In sentencing tasks, the primary objective is to fairly prevent low-risk individuals from reoffending. Therefore, it is essential to control an adequate level of negative predictions for this group and ensure fairness in predictions related to the negative class. Additionally, model

explainability must provide high-quality negative explanations for all individual subgroups to ensure equitable decision-making.

The assessment of fairness in textual analysis and prediction, particularly in the context of recognizing occupation based on textual features, presents unique challenges. While approaches for associating features or sets of features with a specific label often rely on content-based methods — such as thematic analysis — and employ semantic and syntactic aspects for text classification, this task can conceal potential biases rooted in deeper linguistic factors. For instance, sentiment-related nuances, such as the differential treatment of negatively versus positively toned statements, may introduce unintended bias into the predictions (Barsever et al., 2020). Additionally, the specific type of information like social media exhibit biases linked to a limited number of sources, which could lack certain types of information due to the inherent limitations of social media platforms.

For the current study the author employed a hybrid approach to measure the contrastivity of explanations extracted through a post-hoc model-agnostic approach based on Shapley values, by calculating ratios of positive and negative predictions (Lundberg & Lee, 2017). In this work, we assume that all the occupation classes should be predicted fairly for all groups of predictions made, and that the causality of the prediction — demonstrated by a contrast represented in the explanation characteristics — is expected to reflect the fairness of the model's outcomes. Thus, every individual text in the dataset should be predicted as either belonging to a specific occupation (indicated by a predicted label 1 from the sigmoid output) or not belonging to it (indicated by a predicted label 0 from the sigmoid output), based on fair indicators for different groups of predictions.

We structured the analysis of contrastivity based on the nature of Shapley values, represented as positive and negative forces that either increase or decrease the prediction, pushing it toward the positive or negative class (in the case of occupation classification that means — for or against belonging this individual to a specific occupation). This allows us to use the explanation opposite to the predicted outcome, extracted from Shapley values, as a "foil" — the event that did not occur (Miller, 2018, p. 16) — which serves as the background and makes the metric clearly illustrate the causality in the model's prediction explanation.

To ensure a continuous transition for explanations contrastivity values, we invert the numerator and denominator to achieve a coefficient where a value of 1 indicates perfect balance, values less than 1 indicate increasing imbalance as they approach zero, and values close to zero represent maximum imbalance. By doing this, we avoid disjoint or discontinuous ranges that could otherwise make analyzing and interpreting contrastivity more challenging. We compute the contrastivity ratio (contrast) using a normalized formulation, adding 1 to the denominator to avoid division by zero.

Let P denote the sum of all positive values in the explanation, and N represent the sum of the absolute values of all negative contributions. These are computed as 4.1 and 4.2

$$P = \sum_{v_i > 0} v_i, \qquad \text{(sum of all positive values in the explanation)} \tag{4.1}$$

$$N = -\sum_{v_i < 0} v_i, \qquad \text{(sum of absolute values of negative contributions)} \tag{4.2}$$

The contrastivity ratio is then defined by

$$\text{Contrast} = \begin{cases} \dfrac{P}{1+N}, & \text{if } P > N, \\ \dfrac{N}{1+P}, & \text{if } N \geq P. \end{cases} \tag{4.3}$$

This method ensures a unified framework for evaluating contrastive explanations across the entire range of possible values, allowing for a coherent and consistent interpretation. See Listing C.1 in Appendix C. After calculating the SHAP ratios, we examine their distribution for further analysis (Figure 4.2).

The histogram of SHAP-ratio values shows a strong concentration near the upper end of the scale, particularly between 0.8 and 1.0. This pattern indicates that, for most predictions, the model produces highly contrastive explanations, for instance - the explanations clearly distinguish the predicted class from alternatives. There are relatively few cases with low or moderate contrastivity (below 0.7), suggesting the model generally provides confident and distinctive explanations across a wide range of occupations.

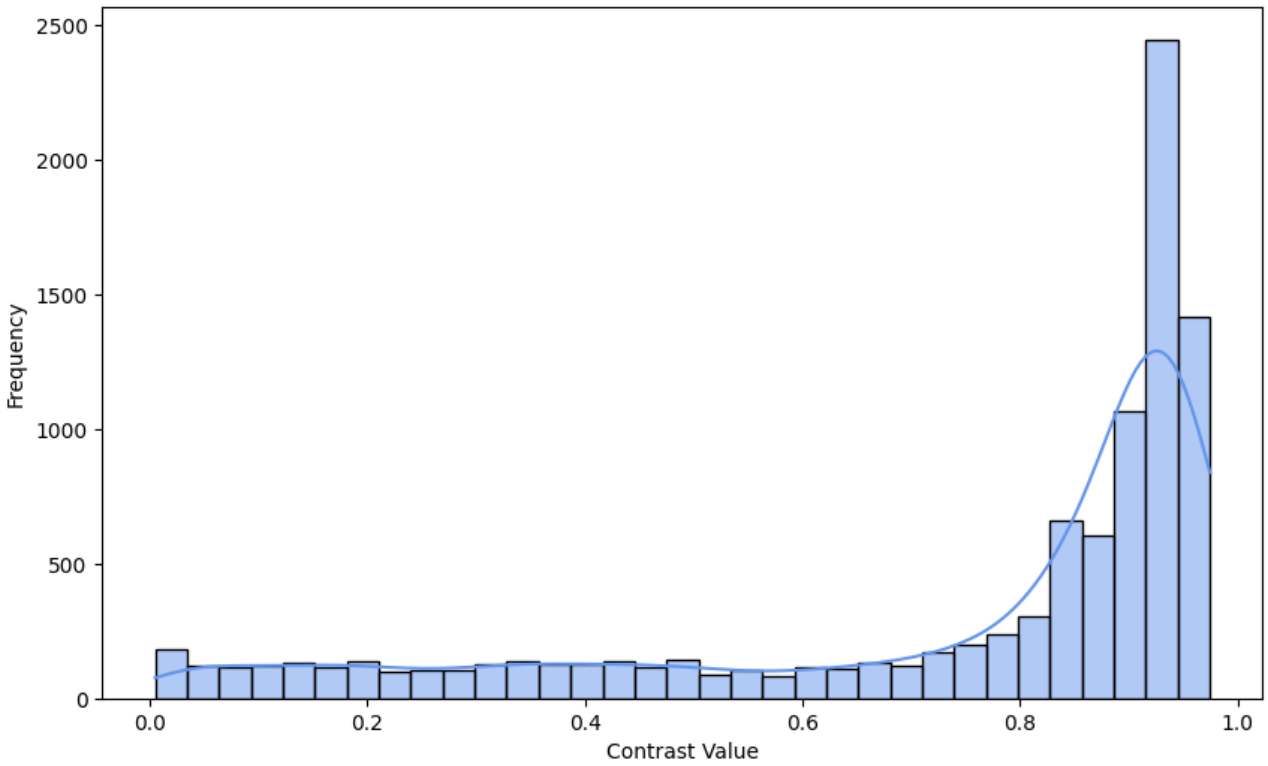

Figure 4.2: Distribution of contrastivity ratios across predictions. Most mass lies between 0.8 and 1.0, indicating highly contrastive explanations.

4.2.3 Exploring contrastivity of explanations for male and female individuals

The dataset used for building our model contains 28 distinct occupations. To ensure a like-for-like comparison of the contrastivity ratio, we compute explanation contrastivity separately by gender for each occupation, adopting a contrastive-explanation perspective (Miller, 2018) and using SHAP-based attributions to obtain the underlying positive/negative contributions (Lundberg & Lee, 2017). We group all contrastivity values by occupation and create two subsamples by binary gender (1 = female, 0 = male). We first examine the empirical distributions to understand their form and visually inspect the spread of values.

Figures B.1 and B.2 illustrate the distributions of explanation contrastivity for predictions made for male and female individuals in two exemplar occupations—*accountant* and *architect*—selected from the full set of 28 classes. The remaining 26 occupations and their contrastivity distributions are provided in Appendix B.

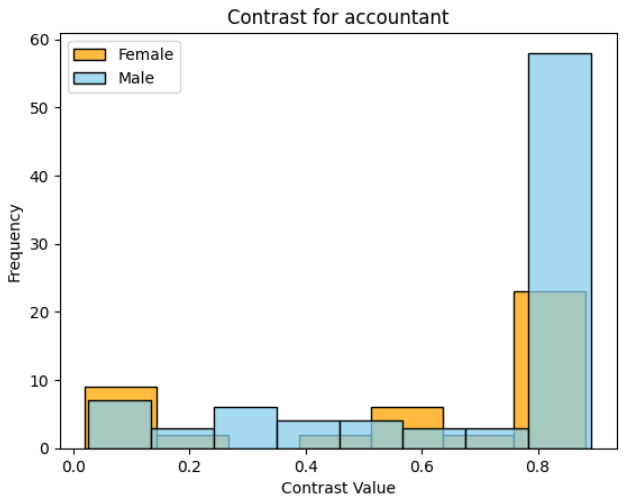

Figure 4.3: The distribution of explanation contrastivity for an accountant.

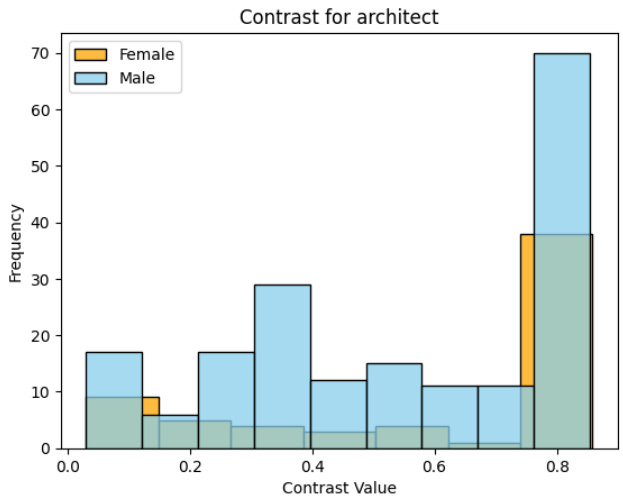

Figure 4.4: The distribution of explanation contrastivity for an architect.

For the accountant occupation class both distributions are right-skewed and majority of contrastivity values clustered between 0.8 and 1.0 for both genders. There is a higher number of high contrastivity values for males, as seen in the taller blue bars at the upper end of the scale. Female individuals show a slightly wider spread across bins, with some additional frequencies at mid-range contrastivity (e.g., 0.6), compared to males. Overalls, the contrastivity tends to be high for both male and female individuals, but the distribution for males is more concentrated at the top end of the scale.

The contrastivity for an architect occupation, separated by genders demonstrate the similar patterns for both genders, having a prominent peak at the highest contrast value range (0.8–1.0). The male distribution is more spread out across lower contrast value bins (e.g., 0.0-0.6), while the female distribution is concentrated primarily at the highest bin (0.8-1.0).

At this stage of the study, we perform a Student's t-test to compare the means of explanation contrastivity values for males and females within each of the 28 professions. This statistical test allows us to determine whether the contrastivity of explanations for male and female individuals differs significantly for each occupation.

Table A.3 in Appendix A summarizes the results of Student's t-tests comparing explanation contrastivity between genders across 28 profession classes. For most professions—including accountant, architect, attorney, and dentist—there is no statistically significant difference

in explanation contrastivity between male and female samples. However, significant gender differences were observed in a few professions: model, nurse, pastor, physician, rapper, and teacher. In these cases, the t-test indicated either higher or lower mean contrastivity for one gender relative to the other, as reflected by the sign of the t-statistic. Overall, these results suggest that while explanation contrastivity is generally similar across genders for most professions, notable disparities exist in specific occupations.

4.3 Ensuring fairness of the BERT implementation for occupation classification

The dataset used to train the occupation–classification model exhibits marked class imbalance, with some occupations far more prevalent than others (Haibo He et al., 2008). Each text instance also includes a binary gender attribute. The frequency of texts for a given occupation by gender reflects real–world occupational segregation, so inherited biases from the environment may influence both model performance and fairness (De-Arteaga et al., 2019, p. 3). Such demographic and class imbalances motivate explicit fairness analysis and monitoring (Barocas et al., 2023).

We visualize the gender composition of each occupation in Figure 4.5 (training set) and Figure 4.6 (test set).

As shown in Figures 4.5 and 4.6, the gender balance in both the training and test datasets is notably uneven across professions, with many fields showing pronounced gender dominance. Professions such as *dietitian*, *model*, *teacher*, and *interior designer* are heavily female-dominated, with female representation often exceeding 80% in both sets. In contrast, careers like *rapper*, *software engineer*, *surgeon*, *filmmaker*, and *DJ* are strongly male-dominated, sometimes accounting for more than 85–90% of the biographies. Only a small number of occupations, such as *journalist*, *painter*, and *paralegal*, exhibit a relatively balanced gender distribution. Overall, these patterns reveal significant gender disparities that are consistent across both datasets, reflecting real-world occupational trends and highlighting potential sources of bias in the data.

These imbalances in the training data—and mirrored in the test data—are reflected in the model's classification report (see Table A.4 in Appendix A). Under-representation of certain

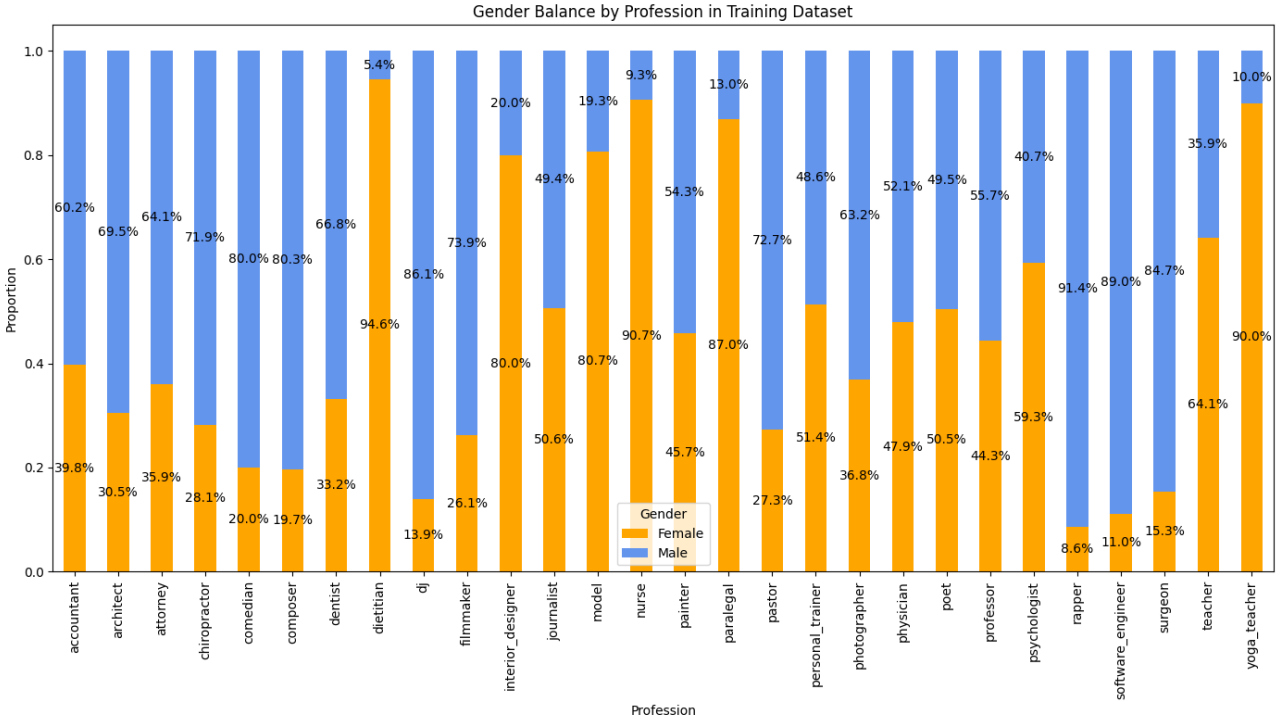

Figure 4.5: Gender balance by occupation (training set).

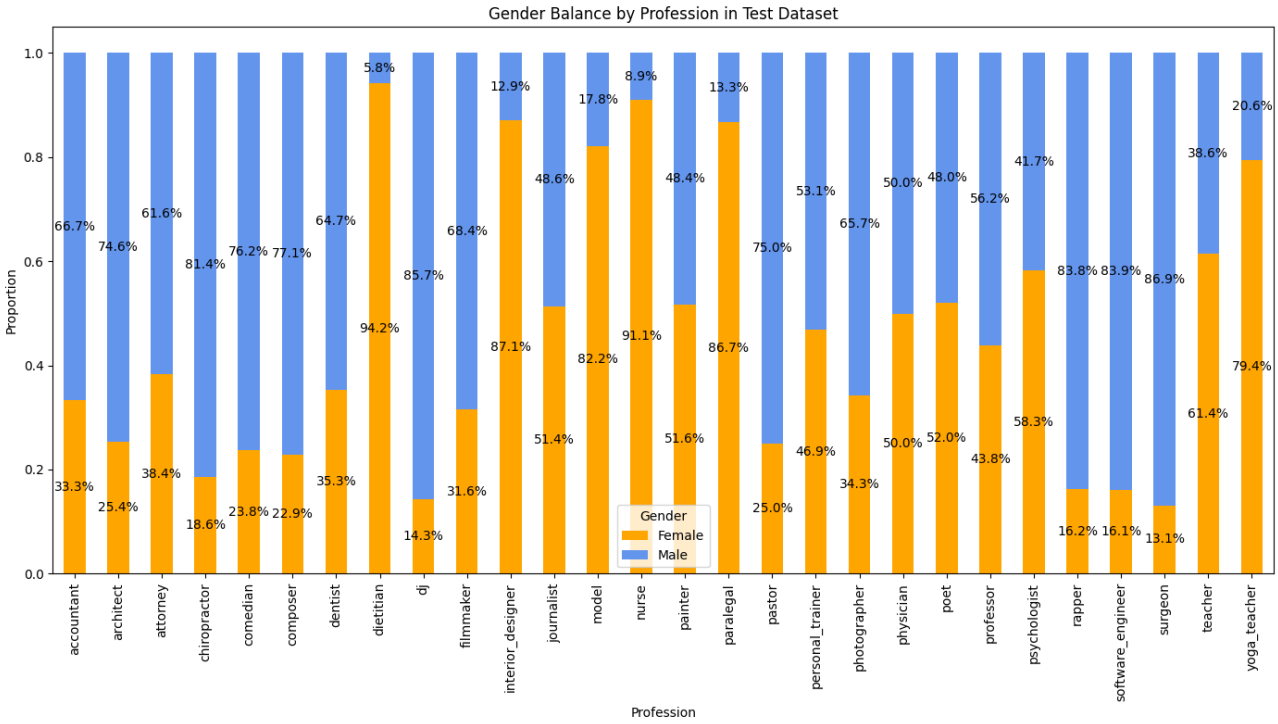

Figure 4.6: Gender balance by occupation (test set).

genders within specific professions often leads to lower accuracy for those gender groups. Imbalances in the test set affect overall model accuracy, while imbalances in the training set may limit the diversity of patterns the model can learn for underrepresented groups. This results in a less comprehensive understanding of the full range of linguistic features associated with certain genders in particular professions, potentially leading to biased or incomplete predictions for those groups.

The BERT model's classification report for female individuals demonstrates an imbalance in the training dataset (Table A.4 in Appendix A) compared to males (Table A.5 in Appendix A). The underrepresentation of a profession relative to gender shows that some professions are predominantly male, while others are female-dominated. For instance, the *chiropractor*, *comedian*, *composer*, *dj*, *pastor*, *rapper*, *software_engineer*, and *surgeon* classes have a high number of men's biographies, whereas the *dietitian*, *interior_designer*, *model*, *nurse*, *paralegal*, and *yoga_teacher* professions are mostly associated with women (Table A.6 in Appendix A). While the BERT model achieves an overall profession-recognition accuracy of 81%—with similar accuracies for occupation recognition for males and females (81% and 82%, respectively)—a closer look at the per-class scores reveals differences in the model's ability to predict certain outcomes for specific individuals.

Table A.6 in Appendix A demonstrates the list of professions for which the male and female exemplar sets differ by at least 30%. Because many classes are imbalanced, we also examine the accuracy of the model's outcomes for each group, following the procedure in (Zhao et al., 2022). Professions that exhibit lower accuracy for a particular gender are highlighted in gray; the remaining imbalanced professions do not show a notable accuracy difference.

We identify professions that exhibit a difference in accuracy between genders. However, such differences do not always coincide with gender imbalance in sample counts. We therefore highlight occupations that, despite comparable numbers of male and female instances, still show accuracy gaps (see Table A.7 in Appendix A). For these cases, we hypothesize that the disparity may be related to features specifically associated with biographies written about females; across these balanced groups, accuracy for females tends to prevail.

4.4 The model inconsistences revealed by the explanation contrastivity

Aligning the findings on explanation contrastivity between the two demographic groups (see Table A.3 in Appendix A) with the observed differences in accuracy and the underlying data imbalances, we find that groups exhibiting both types of inconsistencies—disparities in model predictions and imbalanced training data—also show differences in explanation contrastivity. For completeness, Table A.8 in Appendix A consolidates these results and illustrates how the contrastivity metric behaves in the presence of model inconsistencies.

The pastor, rapper, and model occupation classes, while having data imbalance reflected in accuracy differences, also show lower explanation contrastivity for one demographic group compared to another (highlighted with grey). This is confirmed by the t-test that compares the means of the explanation contrastivity ratio for the two groups and can be explained as the ability to justify the belonging of exemplars to a specific class. The explanations for these groups are less solid, and the contrastivity metrics may be a good instrument for measuring the quality of each prediction to ensure its fairness.

The two cases where the explanation contrastivity metrics revealed data imbalance or accuracy difference separately were the nurse and teacher professions, respectively (highlighted with grey). While these two professions demonstrate those inconsistencies separately, they have enough exemplars for the t-test to reveal a significant difference in the means for both groups.

The opposite situation occurs for professions that have a large imbalance in the data (interior designer, yoga teacher) or an accuracy difference (DJ and dietitian), but the difference in contrastivity does not reach the significance threshold for these groups (highlighted with red). The greatest lack of items among all the classes in the test dataset might be the reason for not reaching test significance, which would allow the difference in contrastivity to be revealed. Still, the phenomenon of contrastivity disparity may be observable in the explanation ratio metric and can be practically applicable to reveal model inconsistencies through the model's explanations.

The accountant and surgeon professions, while not reaching the 5% significance threshold, exhibit differences in accuracy of 19% and 17%, respectively, with t-test significance values of 0.11 and 0.08 (highlighted with light orange). These values are quite close to the conventional significance level and may suggest that, for this specific case, a 10% significance threshold might be more applicable for observing inconsistencies in the model's predictions through explanation contrastivity metrics.

The only profession for which the contrastivity of explanation doesn't demonstrate similar pattern is the physician (light blue). While the test set is balanced and the accuracy for both groups is nearly the same, the t-test revealed a significant difference in explanation contrastivity, which may indicate a lack of reasoning for a specific group. Both the male and female datasets contain a similar number of examples and show only a very small accuracy difference (3%). This may suggest that, in cases where groups are highly balanced, the contrastivity metric becomes too sensitive to minor accuracy differences. Therefore, based on findings from the other professions, we can state that although explanation contrastivity may sometimes signal a problem where none exists (producing a false positive result), the issues for most classes were correctly identified.

To sum up, the ability of the metric to distinguish between fair and unfair predictions, show set imbalances, and statistical disparity is quite high — the metric clearly distinguishes between exemplars that have inconsistencies and those that do not. 66.7% of class exemplars with accuracy bias were correctly identified. In models with enough exemplars in the set (whereas the data is imbalanced), the metric's ability to show bias is expected to be more effective.

However, in around 33% of cases, the metric does not show statistical evidence for recognizing groups that are inconsistent but contain very few items, although there may still be some practical implication. Additionally, it does not strongly represent class imbalance when accuracy rates across groups are similar. There is also a rare case (around 1%) where the model classifies a correct prediction as having an inconsistent explanation.

This is the end of Chapter IV, where we trained a model for textual classification of occupation and evaluated the effectiveness of our explanation metrics. In summary, the metric demonstrated

strong capability to differentiate fair from unfair predictions, highlight set-level imbalances, and surface statistical disparities. It clearly separated instances exhibiting inconsistencies from those that did not, correctly flagging 66.7% of class exemplars with accuracy-related bias. We anticipate even stronger bias detection when models are evaluated on classes with a larger number of exemplars, particularly in imbalanced datasets, where the metric's sensitivity is expected to improve.

In Chapter V, we plan the system architecture for achieving model transparency by implementing cognitive engineering approaches to enable a multi level modeling of system behavior. To reach the main goal of transparency, we will combine model explanations — including Explanation Contrastivity — with assurance of balance at various stages of model training and data preparation. This will support clear identification and categorization of events arising during data preprocessing and model training, promote bias awareness, and enhance explainability of the model's outcomes. The quintessence of these metrics, combined with the multi level approach from CWA (Rasmussen, 1985), forms a resilient, transparent system model architecture.

Chapter 5

Analysing, Identifying and modeling of key functional relationship

In this chapter, we analyze the domain of AI model development and design the architecture of an AI system in accordance with contemporary requirements for system monitoring, data quality assurance, bias mitigation, and the prevention of discrimination (U.S. Government Accountability Office, 2021, p. 30). We conduct this analysis through CWA (Cognitive Work Analysis) (Vicente, 1999) approach implementation for analysing the key values and priorities of the domain. Within the WDA, we performed an Abstraction–Decomposition Analysis that integrates a means–ends assessment with a part–whole decomposition. This produces an Abstraction Hierarchy—ranging from functional purpose, through abstract and generalized functions, to physical function and physical form—and, crucially, reveals the key functional relationships and dependencies that shape system behavior.

The system integrates two fundamental concepts. Technically, it is a complex architecture that combines probabilistic decision-making derived from the use of a black-box model with deterministic components tailored to specific functions. These deterministic elements are low-level engineering techniques that require specialized expertise to understand how the system operates.

On the user side, the target operator is a specialist in human resource management whose responsibilities include user profiling and the adjustment of recommendations based on textual inputs, such as biographies submitted on job-search platforms. The principal difficulty in building systems that support transparency and explainability lies in translating complex quantitative

methods into human-comprehensible representations that allow domain experts to interpret results accurately. In addition, it is essential that the system provide users with the necessary tools to carry out their tasks—specifically, the evaluation of model outputs in terms of fairness and the assessment of the adequacy of the model's reasoning in automated decision-making.

5.1 Applying the WDA and CWA algorithms to analyse and identify the relationships in the algorithms for HR management domain

In this section we foreground Cognitive Work Analysis (CWA) as the primary lens for designing decision support in complex, correspondence-driven domains. CWA begins with the constraints that the environment imposes on action, modeling goals, functions, and affordances in the work domain before specifying cognitive requirements (Vicente, 1999). This ecological emphasis is crucial for sociotechnical systems—such as HR decision pipelines—where dynamic physical and social realities shape what operators can do and how decisions should be governed.

Building on the system aims outlined earlier—system monitoring, data quality assurance, bias mitigation, and the prevention of discrimination (U.S. Government Accountability Office, 2021, p. 30)—we use CWA to structure how transparency is engineered into the AI pipeline. CWA informs work-domain modeling to surface environmental and organizational constraints relevant to fair and lawful decisions; it clarifies control tasks for reviewing model outputs; and it aligns HR specialists' workflows with the system's architecture that combines probabilistic, black-box decision-making with deterministic components. In tandem with Work Domain Analysis (WDA)—hierarchical and abstraction–decomposition frameworks—CWA shapes the transformation of quantitative outputs into human-interpretable evidence, ensuring that fairness diagnostics and the reasoning behind automated decisions can be accurately interpreted and acted upon by domain experts.

Initially, we outline a system architecture designed to enable transparency. The task of recognizing and classifying texts by profession is multi-phase and may comprise numerous sub-tasks, each capable of influencing the final outcome. Automatic decision-making based on a black-box model such as BERT entails several limitations, including, as documented in prior

work, spurious associations between personal pronouns and favorable predictions for particular professions (De-Arteaga et al., 2019). In practice, the system is expected to operate without bias, and predictions of professional affiliation should not be affected by any demographic attributes, such as gender. Professional labels and downstream conclusions should be grounded in descriptions of skills and domain-specific jargon characteristic of the relevant field. Regardless of gender, every individual should be fairly characterized as a representative of their profession based on objective indicators, free from gender bias.

Many professions have historically been regarded as predominantly male; in the present dataset, *pastor* is one such example. Conversely, certain professions have been viewed as predominantly female, such as *nurse* in our case. Today, however, gender balance across professions is being restored, and many sectors support broader trends toward gender, racial, and age diversity. It is essential that these societal imperatives be reflected in systems for automatic decision-making—especially AI-based systems—because their latent layers can obscure the reasons behind particular decisions.

To design our system in accordance with these requirements, we propose an interaction scheme aligned with the human-in-the-loop paradigm, in which a human participates directly in the decision process. After the system classifies texts and, for each biography, predicts the corresponding profession, all results are reviewed by a human operator. Our experimental dataset contains 10,000 records. Contemporary high-throughput recruitment and job-hunting platforms process vast numbers of biographies, résumés, and related documents, and similar profession-classification decisions are routinely derived from textual information at scale. Automating the classification step significantly accelerates data processing and decision-making. However, manual verification of AI outputs can be time-consuming, particularly in large-scale systems. To mitigate operator overload, it is therefore necessary to conduct preliminary analysis and aggregate results in ways that support efficient human review.

At the first phase of CWA we create a work-domain representation using the abstraction–decomposition space, a tool specifically designed to build such models. The abstraction–decomposition space frames a work-domain representation that is distinct from task descriptions: like a map versus turn-by-turn directions, it captures the invariant structure of the controlled system—its purposes, values, functions, processes, and components—independent of any specific

actor, goal, or interface, whereas tasks specify what actions to perform and how to achieve particular ends. Although the decomposition hierarchy is comparatively intuitive, the abstraction hierarchy is harder to interpret (Vicente, 1999, p. 156).

A parable of the ant on the beach, proposed by Gibson (2014), as interpreted in ecological design discourse underscores that observed trajectories reflect not only the actor's internal mechanisms but also the constraints imposed by the environment; understanding behavior therefore requires a stable description of those environmental constraints. Applied to AI prediction of occupation in HR management, the abstraction–decomposition space functions as the domain representation—the "beach"—that shapes and constrains model behavior across the lifecycle: it informs data preprocessing by articulating permissible inputs, privacy and legal constraints, and required invariances; it conditions feature-importance interpretation by linking low-level evidence to higher-level competencies and by preventing discriminatory proxy use; it establishes the salient group structure for fairness assessment by anchoring performance metrics and parity checks in legally and organizationally meaningful categories; and it guides explanation extraction by aligning concrete signals (features, tokens, embeddings) with mid-level functions (gender-based grouping using different attributes of aggregation) and high-level purposes (lawful, non-discriminatory decision support). We illustrate the work-domain representation for two tasks modeled within the system architecture—explanation extraction and analysis, and fairness assessment—in Figure 5.1 and Figure 5.2.

Because these constraints persist across training, data preprocessing, and operations, they also direct choices such as subgroup-aware threshold calibration, fairness-regularized objectives, and monitoring of feature influence. In this way, the "ant's path" - model outputs and operator actions - can be properly understood and governed through the structured description of the "beach" - data items, classes, explanations, legal norms, organizational values, and system components - provided by the abstraction - decomposition representation.

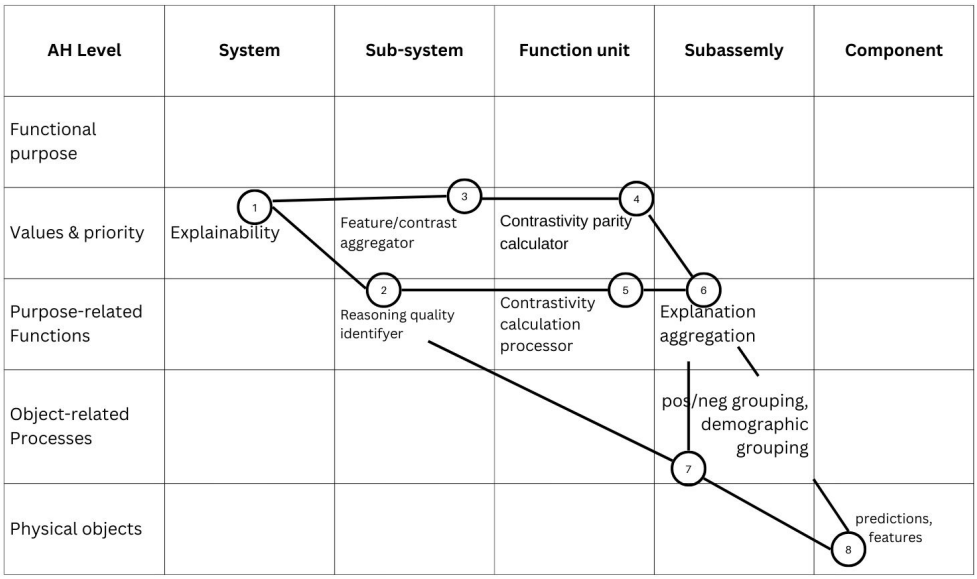

Figure 5.1: A problem solving trajectory of reaching explainabiliy of model's outcome

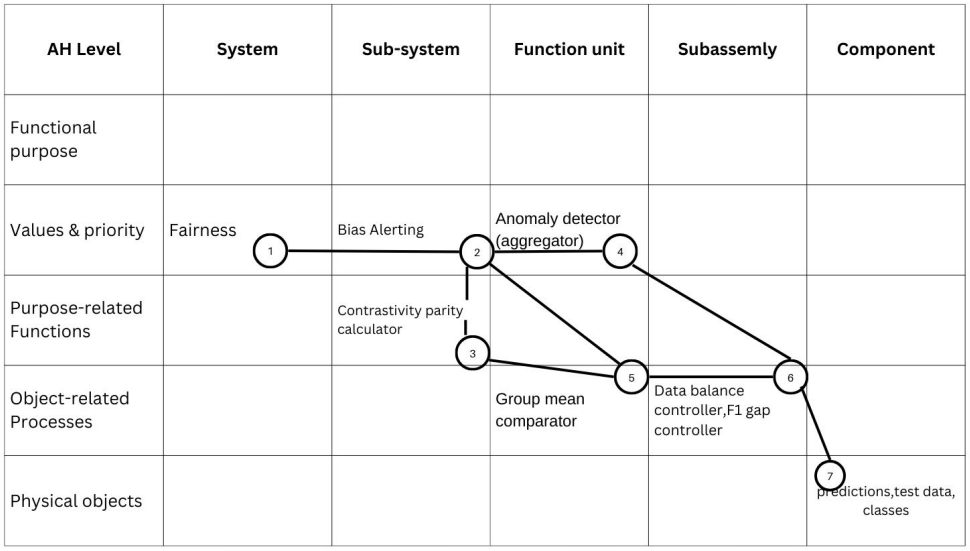

Figure 5.2: A problem solving trajectory of reaching fairness of model's outcome

5.2 WDA for goal-directed problem solving in transparent AI-system

In this section, we explain how the Abstraction Hierarchy (AH) offers a psychologically relevant way to represent work domains. It is conceptualized as a hierarchy of individual elements structured across multiple levels of abstraction (Rasmussen, 1985). From a psychological perspective, a key property of the AH is that higher levels are less detailed than lower levels. Shifting a representation from a low, highly detailed level to a higher, coarser level reduces apparent complexity and thus provides a mechanism for coping with it. In this way, the AH allows resource-bounded actors to manage work domains that would be unmanageable if they had to be apprehended in full detail all at once (Vicente, 1999, p. 173). This advantage—which is unique and psychologically important about the AH—is that it is explicitly purpose-oriented. The levels of the hierarchy are connected by a structural means–end relation, which provides a powerful source of constraint that actors can exploit. Problem solving can proceed by starting at a high level of abstraction, selecting the lower-level function relevant to the current situation, and then focusing on the subtree connected to that function of interest. This approach is efficient because all work-domain elements irrelevant to the current function can be ignored. Consequently, an AH representation supports goal-directed problem solving in a directed and computationally economical manner (Vicente, 1999, p. 174).

We describe the process of AI-model development and operation using the AH, structuring the process from lower-level objects that ML engineers operate to higher, more abstract levels, toward reaching the main system goal—the transparency of the AI system. We start the section by observing strategies for reaching AI-model fairness assessment. The strategies of fairness assessment in AI models are typically categorized as either *group fairness*, which examines the consistency of model outcomes across different demographic categories, or *individual fairness*, which focuses on whether similar individuals receive similar predictions. In the current system design, an enhanced technique of fairness assessment is represented. We present a hybrid approach in which lower-level data objects (e.g., group counts, class labels, demographic statistics) are systematically clustered and transformed by higher-level functions—such as aggregators, balance metrics, or parity checks—finalizing in composite, higher-level representations that map directly

to the system's strategic objectives. The abstraction hierarchy of the system is presented in Figure 5.3. To illustrate how specific nodes relate to the process of achieving fairness through the system-interface design, we have highlighted these nodes in red.

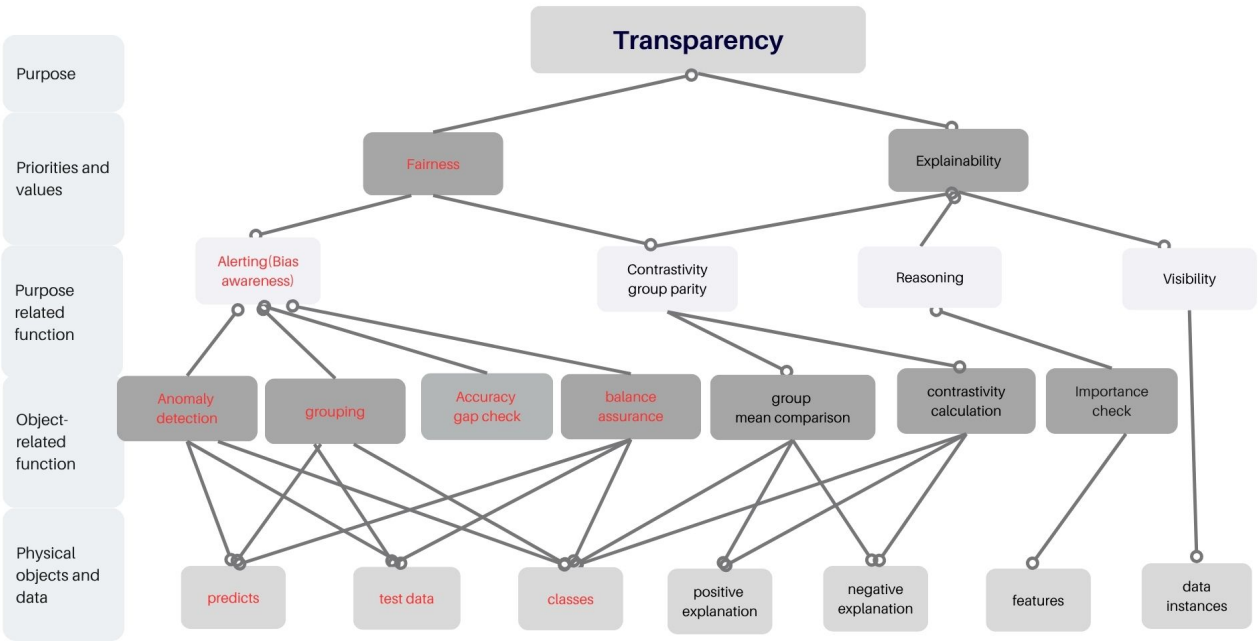

Figure 5.3: Abstraction hierarchy of the system.

This layered structure enables an operator not only to observe the system's foundational components but also to interpret their transformation into actionable, goal-oriented outcomes. In the context of fairness, the operator examines individual data points (at the lowest level), which are then grouped and evaluated using fairness metrics and parity-assessment functions at higher levels of abstraction (Vicente, 1999). The synthesis of these metrics provides an interpretable, high-level assessment of system fairness.

From this perspective, *balance assurance* (ensuring equitable representation or treatment across relevant groups) and *parity assessment* (quantifying and ensuring the equivalence of outcomes or opportunities) are elevated as the central, higher-level functions of the fairness-assessment process (Barocas et al., 2023; Zhao et al., 2022). These functions serve as critical mechanisms to guarantee that the AI/ML system progresses toward its primary goal: enhancing model transparency (U.S. Government Accountability Office, 2021).

Thus, our proposed approach positions fairness not as an isolated goal but as an integral aspect of system transparency, achieved through the systematic aggregation of lower-level data and operations into meaningful, higher-level parity and balance assessments (Vicente, 1999).

We present a bias-alerting mechanism structured as a high-level function within the system, synthesizing groups of low-level data objects by applying a series of transformations and equations. The process enables system operators to observe and understand how low-level imbalances propagate—and are ultimately surfaced—through transparent metrics and visual representations on the system interface, in direct alignment with the system priority of reaching fairness as well as the core transparency goal (U.S. Government Accountability Office, 2021). This mechanism is implemented as a stepwise transformation, starting from lower-level object manipulation and progressing to the highest-level abstraction function of the system (Rasmussen, 1985; Vicente, 1999).

We approach model explainability from a cognitive-engineering perspective, positioning explainability not as an isolated feature but as an integral part of the overall system design (Vicente, 1999). The system is structured as an abstraction hierarchy, where each component—from fundamental data instances to the highest-level reasoning functions—serves a distinct purpose in enabling human operators to interpret, scrutinize, and ultimately trust the system's decisions (Rasmussen, 1985).

By leveraging cognitive-engineering principles, the explainability mechanism is embedded across all abstraction layers, ensuring that model outputs are not only accurate but also understandable and actionable for diverse stakeholders. This hierarchical structure supports operator reasoning, fosters transparency, and provides direct pathways from detailed data and feature analysis up to comprehensive, human-centered explanations (Vicente, 1999).

As illustrated in Figure 5.4, the explainability branch of the abstraction hierarchy encompasses core functions such as reasoning, contrastivity calculation, importance assessment, and the generation of positive and negative explanations—each highlighted to elucidate their contribution to the system's overarching transparency objectives.

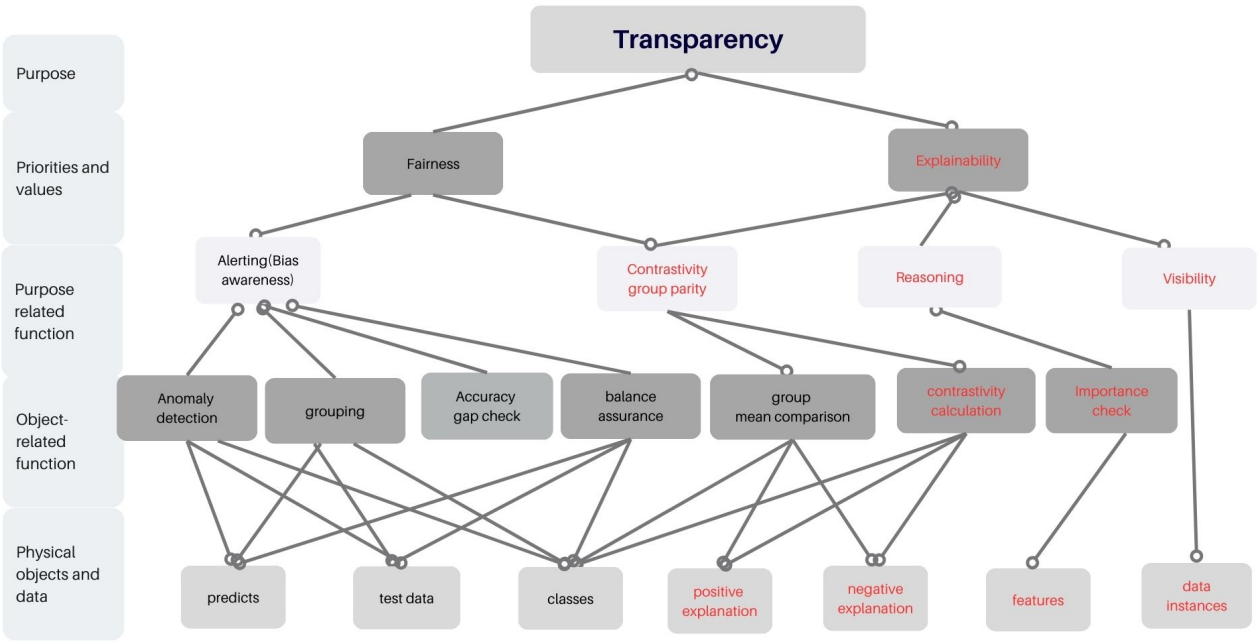

Figure 5.4: The explainability branch of the system Abstraction Hierarchy

At the highest level, the value of explainability is represented as a strategic priority directly supporting the system's transparency goals. This is realized through the system's commitment to enabling *reasoning*, which is the essential function connecting technical processes to human comprehensibility. Reasoning represents the system's capacity to make its decisions traceable and its logic interpretable—essential for fostering user trust and operational oversight.

Moving one layer down, the purpose-related function of reasoning is transformed into two object-related analytic mechanisms: *contrastivity calculation* and *importance check*. Contrastivity calculation allows operators to explore how variations in input features or conditions can lead to changes in predictions, laying the groundwork for meaningful "what-if" analyses and insight into causal relationships (detailed further in a subsequent section). The importance check, in turn, quantifies the influence of each input feature on a given prediction, providing direct evidence of the internal factors driving the model's conclusions.

Applied to AI in the HR-management domain, we structure model development and operation from low-level data objects to higher-level functions aligned with the system's strategic goal of transparency, treating fairness assessment and explainability as integral branches within the Abstraction Hierarchy rather than isolated features: lower-level counts, labels, and demographic

statistics are aggregated via parity and balance metrics into composite indicators, while reasoning, contrastivity, and feature-importance analyses translate technical evidence into human-interpretable explanations. This technique directly supports the effective visualization design discussed in Chapter VI.

Chapter 6

Designing the visual representation for algorithmic transparency

In this chapter, we design the system's visualization to represent the key functions of the work domain and to develop a pathway toward the overarching goal of making AI outcomes more transparent. To engineer transparency, we employ an Algorithmic Transparency Framework(Hepenstal et al., 2019), which offers significant advantages for analyzing model outputs by enhancing the visibility of reasoning, evidence, goals, and constraints that support the analysis.

6.1 Algorithmic transparency framework implementation

The formal concept analysis allows to understand reasons behind formal mathematical and systematic methods within conceptual (meaningful, interpretational) approach. In the initial work toward the framework development and implementation it was proposed as a framework for interpretability of AI systems that makes the relationship between various system elements visible and helps identify processes mandatory for the reaching the outcome. This is an intent-based framework emphasize the approach of an interpretation that can be decomposed for opportunity of receiving an explanation for each component (Figure 6.1).

In the present implementation, dynamic decomposition of operator intent is not required. Instead, operator intent is distilled from the initial system requirements - namely, the provision of individual and group fairness assessments and individual explanations that yield defensible

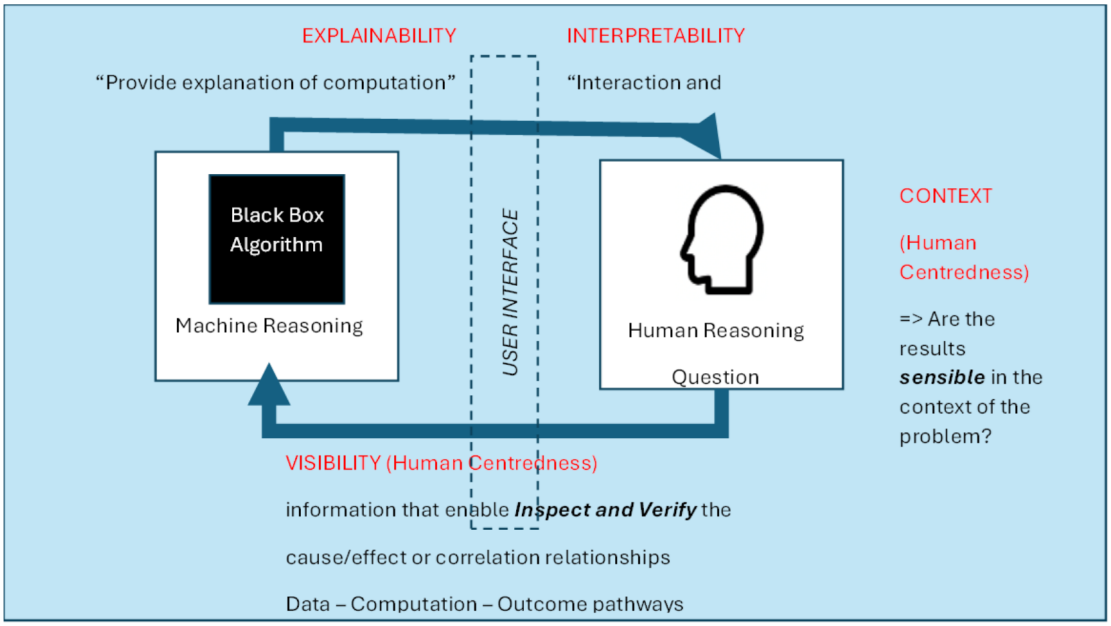

Figure 6.1: Algorithmic Transparency Framework. Adapted:Hepenstal et al. (2019)

insights into the model's reasoning. This focus reflects a broader principle in human-AI interaction: effective information provisioning enables practitioners to assess whether model predictions are sensible and, where appropriate, to challenge them. Within this system, operators are supported with visual representations of model and data elements at higher levels of abstraction to support two primary purposes: assessing the fairness of model outcomes and evaluating the quality of reasoning underlying individual predictions.

To operationalize transparency, we structure fairness analysis as a staged process in which each phase is explicitly linked to the explanations the system must deliver. This ensures that transparency is not treated as a supplement but as the culminating objective of the interaction.

In the first phase, users assess group fairness. The central interface requirement at this stage is to enable clear differentiation between fair and unfair group outcomes and to characterize the specificity and nature of any detected disparities. The interaction logic is grounded in the Algorithmic Transparency Framework, which organizes goals and constraints for recognition by the operator and thereby supports accurate interpretation of model behavior. According to the Formal Concept Analysis approach, we transform the mathematical formalism into a conceptual representation; thus, the model's outcome associated with a specific gender label, produced by a

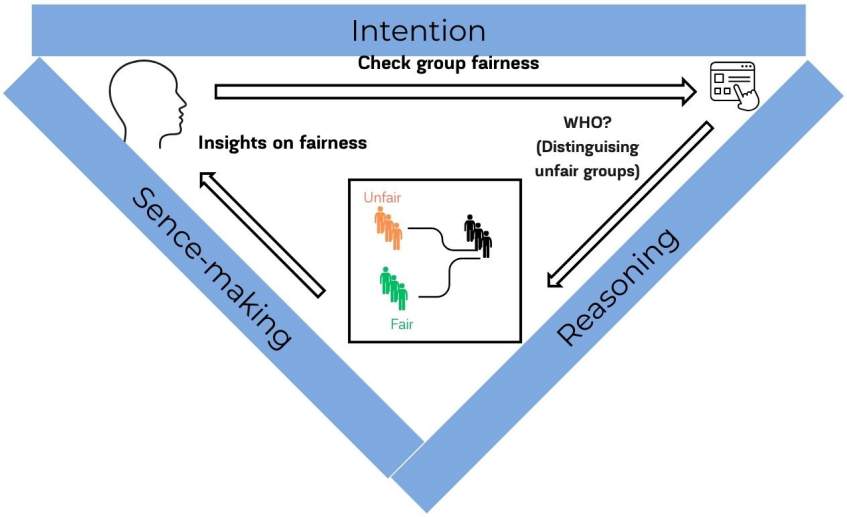

Figure 6.2: Conceptual representation of group fairness assessment function of the HR-management system

group fairness assessment algorithm, is interpreted into a representation that supports a user's understanding in the context of their current knowledge of the model's outcome (Figure 6.2).

In the second phase, users deepen their understanding of the sources of unfairness by examining individual outcomes, associating predictions with textual features, and reviewing the quality of the accompanying explanations. This phase strengthens the evidential link between observed disparities and the model's internal reasoning. As the user forms a specific frame of understanding of group fairness, subsequent intent and interaction with the interface develop a deeper understanding of the reasoning underlying the specific unfairness (Figure 6.3).

In the final phase, users synthesize the insights gained at the group and individual levels to form a coherent judgment of the model's overall behavior. The user interface delivers this synthesis through transparent, resolution-controlled views that connect high-level fairness assessments to their underlying evidential bases. Thus, by incorporating the Algorithmic Transparency Framework, we define three stages for shaping user understanding of the model's outcomes by interpreting results so that users can distinguish different inconsistencies-beginning with group unfairness and data imbalance and culminating in the reasoning behind individual predictions.

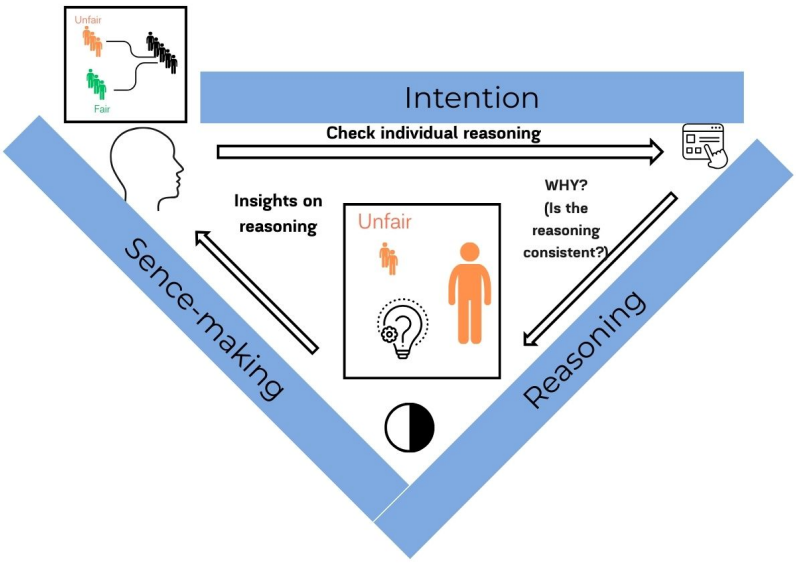

Figure 6.3: Conceptual representation of individual fariness assessment function of the HR-management system

Consistent with the cognitive engineering perspective, our approach treats explainability as a system-level property rather than an isolated feature. The system's abstraction hierarchy maps lower-level objects (e.g., feature contributions, demographic counts, class labels) to higher-level functions (e.g., parity checks, balance metrics, contrastive analyses), culminating in value-oriented goals (transparency, accountability, and contestability). By articulating the means-end relations among the elements of the system, the interface supports expert reasoning about the user-cases: operators can inspect how lower-level imbalances propagate through analytic transformations, verify that detected unfairness is grounded in evidence, and determine whether predictions warrant challenge. In this way, the design aligns the classification of operator intent with the structure of the analysis itself, thereby enhancing the interpretability of system outputs and the efficacy of analyst interaction.

6.2 Designing the Visual Representation: Applying EID principles for AI Transparency

In the following sections, we focus on observing the applicability of Cognitive Engineering approaches in AI model development and investigate the strategies that specialists in this area

utilize to achieve model explainability, fairness, and transparency. We explore the functions, represented by equations, that formulate these processes mathematically. Additionally, we develop a hierarchical representation of the domain that allows us to address the system's main goal by incorporating lower-level objects and presenting them through higher-level abstract functions, mapping them onto the geometry of the user interface. We also discuss the application of the Contrastivity of Explanation metric and its implementation in the user interface to support one of the key system priorities of achieving model explainability and supporting reasoning for the model's outcome. We begin this chapter with a brief introduction to the basics of Ecological Interface Design (EID) principles, which inform user-interface creation, system-architecture planning, and configural display development.

EID is a framework, initially systematized by Burns, for building human–machine interfaces that has demonstrated strong results across a wide range of domains and display types (Burns et al., 2000; Naikar, 2017; Vernon et al., 2002; Wong & Gulden, 2017). It also supports heterogeneous information representations (Naikar, 2017, p. 523). EID distinguishes between two types of display units—individual and configural—based on the level of information represented. A *configural* display enables an operator to perceive system elements at a global level (not only locally), thereby supporting EID's core objective of enhancing system safety by enabling adaptive behavior and mitigation of unanticipated events (Vernon et al., 2002, pp. 223–224). This benefit stems from organizing elements at multiple levels of abstraction—aggregated according to the abstraction–decomposition space—so that groups of lower-level elements are rendered as higher-level visual forms that convey overall performance (Wong & Gulden, 2017, p. 2).

A key aspect of EID is that it helps operators monitor system state and perceive which actions can be taken to achieve system goals. Sets of possible actions associated with system invariants are manifested in the interface as *affordances*, an important consideration when mapping functional relationships onto a two-dimensional geometry (Vernon et al., 2002, p. 224). Another central feature is EID's explicit treatment of *boundaries and constraints*, which clarifies what is desirable versus what must be avoided. By making operational limits—mechanical/physical limits and safety boundaries—perceptible, constraint-based design directly supports decision making and problem solving (Vernon et al., 2002, pp. 223–225). Additional boundaries tied to

domain practices or human constraints should likewise be made explicit (Tieu & Naikar, 2022, p. 3).

A contemporary extension to human–AI systems emphasizes *self-organization*: when conditions change, participants adjust organizational structures rather than follow rigid procedures. Actors rely on fundamental behavioral guidelines to respond effectively to local conditions, allowing new structures to emerge as needed. This capacity for self-organization is essential for successful adaptation in changing environments (Naikar et al., 2023, p. 1667).

Previously, we articulated a five-level abstraction that ensures model transparency through the user interface—from low-level data objects to high-level, value-oriented goals. In this chapter, we instantiate that level-based approach by beginning at Level 2 with balance checks over classes and demographic groups using Statistical Parity (SP) and the Imbalance Coefficient metrics, aggregating evidence at Level 3 into fairness diagnostics, combining these at Level 4 with explanation structures, and culminating at Level 5 in value-oriented judgments about the outcome's fairness or explainability, aligned with an Algorithmic Transparency Framework. This progression preserves traceability from raw distributions to strategic assessments, enabling users to connect observed disparities to the reasoning behind individual predictions.

6.3 Semantic mapping

Planning the further process of visual representation for AI-system operator support, we use fundamental EID principles originally developed to support decision making in complex systems. In contrast to user-centric approaches—aimed at usability goals derived from user characteristics—EID is a constraint-based approach grounded in the environment (Tieu & Naikar, 2022, p. 3). The processes performed in the system, determined by environmental constraints, form a set of physical and mechanical relations that must be rendered as graphical and visual relations (Vernon et al., 2002, p. 224). The approach to linking visual forms to the meanings derived from the represented domain, as articulated by Bennett and Flach, argues that "what the display looks like must be considered in the context of what the display means" (Bennett & Flach, 1992).

Thus, to support interface design at this stage, we use a semantic-mapping approach (Bennett & Flach, 1992) to transform geometric forms on the display according to equations that characterize system behavior. The essential requirement is to depict the specified relationships accurately and to facilitate appropriate transformations of geometric forms (Vernon et al., 2002, p. 224).

6.3.1 Data balance and accuracy difference object-related function definition

The process of generating visual representation begins from the low-level functions by ensuring the balance of data instances across each class or demographic group. For this purposes we impellent two commonly used metric for balance assurance - (i) **SP**(statictical parity) and (ii) **Imbalance Coefficient**. Statistical parity speaks to group fairness and equalizes outcomes across protected and non-protected groups. (Barocas et al., 2023) The concept of statistical parity in fairness metrics is initially described as the effort to "equalize some group-dependent statistical quantity across groups defined by the different settings of A." For example, this might involve requiring that acceptance rates are the same for all groups, which mathematically imposes the constraint represented by equation 6.1 below .

$$P\{Y = 1 \mid A = a\} = P\{Y = 1 \mid A = b\} \quad \text{for all groups } a \text{ and } b. \tag{6.1}$$

Imbalance Coefficient.

Imbalance Coefficient is the approach of calculating the degree of class imbalance, developed specifically to mitigate learning bias in original imbalance data distribution (Haibo He et al., 2008, p. 1322). Authors describe the approach assuming the training dataset D_{tr} which consists of m samples, denoted as $\{(x_i, y_i)\}_{i=1}^{m}$, where each x_i is a data point in an n-dimensional feature space $\mathcal{X}$, and y_i is its corresponding class label with $y_i \in \mathcal{Y} = \{1, -1\}$.

Thus, m_s represents the number of minority class samples, and m_l the number of majority class samples, such that $m_s \leq m_l$ and $m_s + m_l = m$.

The equation 6.2 below represents the procedure of calculating the degree of class imbalance.

$$d = \frac{m_s}{m_l}. \tag{6.2}$$

The implementation of lower-level functions involves the manipulation of various objects at a detailed level. For example, assessing the accuracy gap between demographic groups using statistical parity (SP) metrics requires aligning data distributions with outcomes or predictions. Notably, the initial application of the imbalance coefficient is closely related to a predefined threshold for the maximum tolerated class imbalance ratio (Haibo He et al., 2008, p. 1333). At the current stage, no operational thresholds have been applied; this phase is purely diagnostic, capturing foundational statistics to reveal whether classes and outcomes are balanced or if disparities exist between groups.

6.3.2 Event segmentation for purpose-related function definition

With foundational metrics established, the system advances to event inference. Here, each "event" represents a synthesized observation - a function of calculated for any of the metrics from object-related functions combined with system boundaries constraints. A categorization mechanism maps these events into alert states by grouping them, identifying prevalent demographic patterns, and applying equation-based logic.

For clear event categorization, we utilize a two-dimensional space where each point represents an event based on its demographic representation and predictive equality between groups. X-axis represents the gender imbalance coefficient (GIC), defined as the ratio of the number of elements in demographic group 1 to group 2. This axis quantifies the balance, or imbalance, in group participation or presence within the event. Y-axis demonstrate the accuracy difference coefficient (ADC), expressed as the ratio between predictive accuracy for group 1 versus group 2. This dimension captures fairness in predictive performance across groups.

Thresholds (which can be customized for each axis) define a central "fair" region, typically forming a rectangle centered at (1,1)(1,1). Events that fall within this region are classified as fair in both representation and prediction, while those outside demonstrate a significant imbalance or unfairness in at least one dimension. This diagram allows for rapid visual assessment of

where and how events may require further scrutiny, intervention, or improvement to ensure both equitable participation and fair predictive accuracy for all demographic groups.

In the equations 6.3,6.5 below we present the unified formula of representing the gender imbalance for number of participants (Imbalance Coefficient) in each demographic group as well as prediction accuracy difference for each group (Statistical Parity):

Gender Imbalance Coefficient (GIC)

$$\boxed{\text{GIC} = \frac{N_1}{N_2}} \tag{6.3}$$

Where

$$N_1,\ N_2\ =\ \text{number of participants in each demographic group.}$$

Fair if:

$$1 - t_N\ \leq\ \text{GIC}\ \leq\ 1 + t_N. \tag{6.4}$$

t — threshold variable, which defines the acceptable range of balance.

Accuracy Difference Coefficient (ADC)

$$\boxed{\text{ADC} = \frac{A_1}{A_2}} \tag{6.5}$$

Where

$$A_1,\ A_2\ =\ \text{prediction accuracy for each group.}$$

Fair if:

$$1 - t_A\ \leq\ \text{ADC}\ \leq\ 1 + t_A. \tag{6.6}$$

t — threshold variable, which defines the acceptable range of balance.

Each axis can use a different threshold (t_N for population, t_A for accuracy), but the fairness pattern remains identical and symmetric.

By utilizing this methodology, we identify 6 distinct regions where a given point—representing the balance and fairness between two groups based on the calculated metrics—can be scattered. These regions are determined by applying upper and lower threshold bounds to each fairness metric (**GIC** and **ADC**) (6.1).

Table 6.1: Event categories represented by fairness diagram

Event	**Event classification**	**Type of event**
1	Both metrics within threshold bound — balance in group proportion, balance in prediction accuracy across groups	Regular
2	GIC exceeds the upper threshold, ADC within bounds — imbalance in group proportion, but prediction accuracy remains fair	Alerting
3	GIC below the lower threshold, ADC within bounds — underrepresentation of one group, with fair prediction accuracy	Alerting
4	GIC within bounds, ADC exceeds the upper threshold — fair group proportions, but prediction fairness skewed in favor of one group	Alerting
5	GIC within bounds, ADC below the lower threshold — fair group proportions, but accuracy biased against one group	Alerting
6	Both metrics outside their threshold bounds — both representation and prediction accuracy are unfairly distributed	Alerting

This approach enables the visualization of fairness and bias alerts as elements on a two-dimensional interface, making disparities immediately visible to the operator. Events are categorized and mapped in this two-dimensional space, collectively forming an alerting system that supports the goals of bias awareness and fairness assurance.

The specific alerting system allows the operator to clearly understand the type of event by showing its position relative to defined fairness thresholds. All events that fall outside of the "fair zone" are automatically highlighted, prompting further attention. Depending on their location in the interface, specific types of bias can be easily identified and flagged for deeper classification or further analysis, ensuring systematic monitoring and proactive response to fairness concerns.

6.3.3 Identification of the emergent criteria for relevant mapping of visual forms

Building the system interface according to the principles of Ecological Interface Design (EID) involves accounting for emergent aspects that must be identified in order to construct enhanced visualizations. Such visualizations should accurately represent the functional relationships of the domain and support the operator in working effectively within the identified boundaries and constraints of the domain environment. The table below outlines the criteria for interface construction based on EID principles and incorporates several emergent aspects (Table 6.2).

System Property	**Purpose**	**Demands**
Constraints and boundaries	Distinguish what is physically or procedurally possible or desirable from what is impossible or undesirable.	Constraints in human–machine systems should be displayed so they can be seen directly (Vernon et al., 2002, p. 224).
Relevant affordances for actions	Allow the operator to see the current state and determine what actions can safely be taken to achieve the system goal.	Map the relational invariants of the work system onto the interface to create a virtual ecology where relevant affordances are visible (Vernon et al., 2002, p. 224).
Physical system elements	Support reasoning about material and functional configurations.	It is recommended to use literal physical representations (Vernon et al., 2002, p. 225).
Abstract system elements	Support reasoning about higher-level processes and functions.	It is recommended to use abstract representations that reveal functional properties not directly observable (Vernon et al., 2002, p. 225).

Table 6.2: Criteria of the interface built within the EID principles

The task of illustrating the processes performed by an AI model - as well as specific aspects of its operation, such as class balance or feature influence - is inherently complex, as it involves the analysis of multiple stages of data processing (Doshi-Velez & Kim, 2017) (Lou et al., 2012).

While there is no fixed criterion for selecting visual forms or developing new visual elements in the context of EID design (Tieu & Naikar, 2022, p. 3); (Vernon et al., 2002, p. 225), our goal is to employ visualization techniques that enable the segmentation and grouping of display

elements in order to enhance effectiveness and reduce the operator's cognitive load (Vernon et al., 2002, p. 223).

Visualization approaches applied across various domains may differ; however, the strategy of constructing system architectures based on functional relationships to support problem diagnostics and efficiency assessment—within the context of the specific domain's values, goals, and priorities—appears to be common across multiple areas of application. Although the selection of visualization methods has been discussed in detail in previous works (Lemieux et al., 2014; Tieu & Naikar, 2022), this study focuses on exploring the application of techniques proven effective for illustrating complex relational architectures and providing substantial support for operators, specifically through configural displays that enable the inspection of lower-level objects via higher-level functions, in accordance with the values, priorities, and system goals.

In the context of this paper, the AI system under study is conceptualized as a complex system in which understanding model reasoning requires a high level of expertise. The transformation of higher-order functions—such as reasoning and fairness—into interpretable insights is cognitively demanding, as it necessitates a deep understanding of both data structures and algorithms. Moreover, the inherent complexity of AI models, arising from their non-linearity and opaque internal processes, amplifies the challenges associated with visualization and underscores the need for advanced, rigorously validated methods developed in domains of comparable complexity.

Thus, to support the visualization of complexity in AI-driven domains and to ensure the proper representation of functions at various levels while incorporating key criteria for system interface construction, we utilize the Risk Mapper design technique proposed by Wong and Gulden (Wong & Gulden, 2017).

The Risk Mapper technique incorporates strategies that enable visual geometries to highlight distinctive and perceptible patterns or emergent features (Wong & Gulden, 2017) and is grounded in strong theoretical principles of system boundaries and constraints (Vernon et al., 2002, p. 224). Based on the assumption of system invariants (Lemieux et al., 2014; Naikar, 2016, p. 4), this set of visualization approaches enhances model transparency within a broader, value-based framework. This enables the complexity of AI model training to be visually clarified, reducing

the burden of information representation and lowering the cognitive load required to interpret results.

6.4 Risk Mapper dashboard implementation for ensuring fairness of AI prediction

This section presents - through the cognitive engineering framework - an in-depth account of designing and prototyping such dashboards. Emphasis is placed on the concept of system borders (Rasmussen et al., 1994), constraints (Norman, 1986); (Woods, 2016), system invariants (Hollnagel, 2008), as well as cognitive strategies for reducing practitioner workload (Rasmussen, 1983) across skill-, rule-, and knowledge-based processing levels.

The responsible deployment of AI systems, especially those involved in sensitive tasks like occupation prediction, hinges on effective methods for detecting and communicating fairness risks. Recent studies underscore the prevalence of gender imbalances in algorithmic predictions, which can silently perpetuate or amplify social biases (Barocas et al., 2023; Haibo He et al., 2008; Morzhov, 2020). In response, the discipline of cognitive engineering - pioneered by Rasmussen, Pejtersen, and Goodstein Rasmussen et al. (1994) - offers a valuable lens for structuring the design and evaluation of interactive dashboards. The RiskMap, a variant of the clustered scatterplot, is especially useful as it encapsulates both data balance and performance disparity in a way that is aligned with human perceptual and cognitive abilities.

Data preparation: aggregation and metric selection

The effective visualization rests on a foundation of relevant, high-quality metrics. Within the RiskMap, the lower-level functions of balance assurance (male/female sample count for each profession class) and parity assessment (magnitude of performance gap between genders) were selected based on clear precedent from the algorithmic fairness literature (Barocas et al., 2023; Chouldechova, 2016; Haibo He et al., 2008) and support the core system priority of AI model fairness assessment progresses towards its primary goal - model transparency enhancement. These indicators qualify not only model output disparity, but their class-level aggregates that

allow practitioners to discover where imbalances may be hidden on both dataset and prediction level.

As the first step of the data preparation process, we segmented predictions by gender for each occupation and computed the Gender Imbalance Coefficient (GIC) and Accuracy Difference Coefficient (ADC), as discussed in detail in Section 6.3.1, for each subgroup. Similarly to Lemieux et al. (2014) we implement we map invariant relationships as x and y axes of quadrant map. At our adaptation for fairness analysis, the horizontal axis is repurposed to represent dataset imbalance, thereby quantifying the "risk" associated with disproportional gender representation within each occupational class (Figure 6.4).

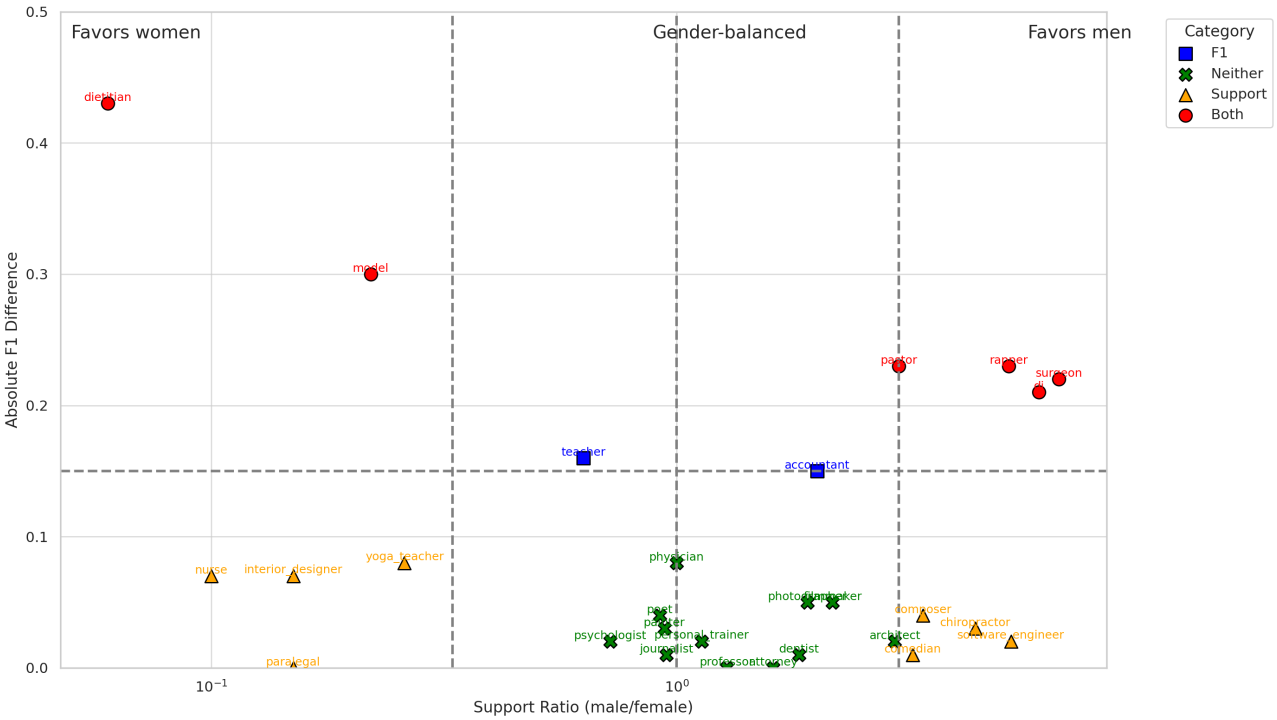

Figure 6.4: 2-dimensional projection of prediction and dataset inconsistencies

Here, the x-axis is centered at 1, denoting a state of perfect gender balance wherein the number of male and female samples is equal for a given class. Occupations positioned at the centroid thus exhibit no gender-based skew — an invariant of demographic parity. Movement to the right or left along this axis indicates increasing imbalance, with the direction specifying the group toward which the imbalance is biased: positive values signal male overrepresentation, while negative values denote female overrepresentation. The degree of displacement from the origin encodes the magnitude of that imbalance, operationalized as the logarithm of the male-to-female

support ratio.

By mapping each class's position with respect to this invariance, the diagram mirrors the original Risk Mapper's capacity to visualize deviation from baseline "functional invariants." In this context, dataset imbalance constitutes a form of risk that may undermine both model equity and interpretability — paralleling how liquidity or market attractiveness inform risk in portfolio analysis in the initial work of Lemieux et al. (2014). This representation thus enables practitioners to systematically identify and interrogate sources of imbalance at a granular, class-specific level, extending core principles of the Risk Mapper approach to the domain of algorithmic fairness.

The y-axis of the Risk Mapper thus encodes, for each occupational class, the degree to which predictive accuracy is different between gender subgroups. Occupations plotted closer to zero on the y-axis manifest high parity in predictive performance across gender categories — functioning as "accuracy invariants" in the fairness landscape, analogous to well-functioning liquidity or market attractiveness in the original Risk Mapper paradigm. By contrast, points further from the horizontal center signal a model's inability to ensure comparable accuracy between groups, irrespective of sample sizes.

This explicit mapping of the ADC on the y-axis allows for a nuanced understanding of algorithmic fairness, distinguishing between mere representation (as charted by the x-axis) and the real-world harm or disparate impact engendered by accuracy gaps.

By interpreting both axes together — the Gender Imbalance Coefficient (GIC) on the x-axis and the Accuracy Difference Coefficient (ADC) on the y-axis — the Risk Mapper provides a nuanced view of fairness across occupational classes. Occupational classes plotted near the origin on the y-axis, where ADC approach zero, and the center of x-axis, where GIC is near 1 demonstrate both equal representation and fair predictive performance for men and women, embodying the ideal of algorithmic fairness. Classes that shift substantially left or right on the x-axis identify the direction and extent of gender imbalance, while those positioned higher on the y-axis reveal occupations in which the model's accuracy differs meaningfully between the two genders. Ultimately, this configuration enables practitioners to detect not just which

occupations exhibit representational bias, but also whether such bias translates into concrete disparities in predictive outcomes.

6.4.1 Principles for Configural Display: Theoretical Underpinnings for Risk Mapper implementation

A distinguishing strength of the Risk Mapper is its rootedness in *configural display* theory. Configural displays are visualizations where abstract quantities (here, fairness metrics) are mapped onto spatial geometries to induce emergent patterns, allowing users to leverage native visual faculties for risk detection (Bennett & Flach, 1992, pp. 170, 183). This method is especially effective for supporting skill-based cognitive operations and for reducing the interpretive burden typically associated with data preparation and model evaluation (Vicente, 1995). In the Risk Mapper context, system borders become visual boundaries that separate quadrants along the axes of support ratio (GIC) and F1 difference (ADC) (Figure 6.5). These quadrants are not arbitrary; rather, they encode thresholds that reflect organizational or legal policy—for instance, tolerable bias levels defined by a maximum permissible F1 difference (ADC) (Barocas et al., 2023). System constraints—rendered by the dashed lines—partition the "safe" region, characterized by balanced representation and low difference, from the more "risky" periphery, where imbalances or higher divergences call attention to potential problems. Thus, the visual structure of the Risk Mapper, supported by a constraint-based approach, provides a well-justified engineering rationale and allows users to distinguish desirable system outcomes from those that are not preferable (Vernon et al., 2002, p. 224).

6.4.2 The implementation of contrastivity metrics using Semantic Mapping

The contrastivity of explanation is the metric that supports the model's reasoning process for each predicted instance and is an aspect that allows the operator to investigate model outcomes from an additional and nuanced perspective. As described earlier in Section 4.2.3, contrastivity of explanation is presented at the object-related level and anchors more abstract constructs, such as reasoning and explainability at the level of system purposes and values, ultimately becoming a foundational element of overall system transparency.

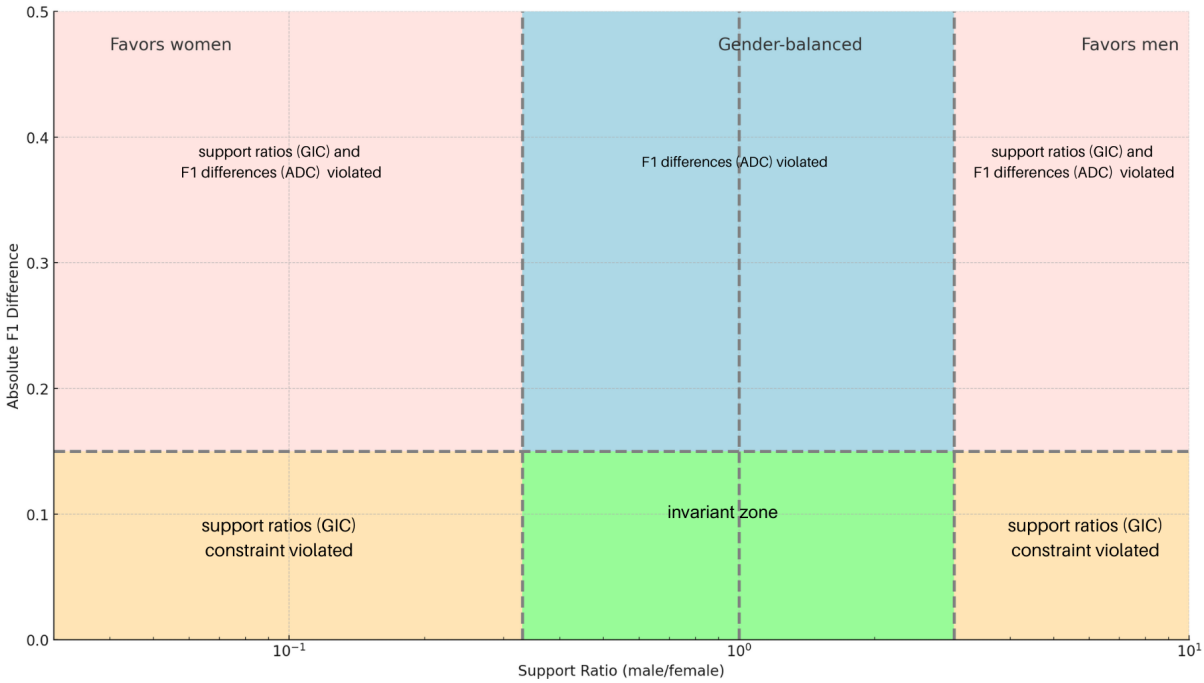

Figure 6.5: Configural display elements to support visual perception for Risk Mapper implementation

While model fairness assessment provides a broad understanding of prediction quality across classes—helping to effectively distinguish specific classes that may require further diagnosis—the measurement of explanation quality and contrastivity serves as an enhanced tool for assessing the depth and clarity of model reasoning. This additional dimension enables operators and domain specialists to perform an enhanced evaluation of model outputs, supporting the research on transparency that emphasized the need to ensure that system operators and users act with accountability and fairness and maintain trust in their systems (Hepenstal, 2023, p. 44). This offers supportive evidence for the robustness of model decisions and thereby supports these essential requirements for trustworthy and responsible system use. Compared to fairness metrics, which typically focus on sensitive groups or aggregated outcomes, explanation contrastivity is calculated for each individual prediction. This granularity grants the metric both an individual and a class-level relevance, positioning explanation contrastivity as a compelling measure that bridges the gap between systemic assessment and case-by-case validation.

Empirical findings presented further in this study indicate that a lack of explanation contrastivity frequently coincides with classes that demonstrate unfairness toward specific demographic groups, reinforcing its value for diagnostic scrutiny. Therefore, contrastivity not only holds significance as a class-level characteristic but also aids in identifying individual instances

where the model's reasoning may be insufficiently decisive. In such cases, low contrastivity can serve as a trigger for further expert analysis, thus supporting a more responsible and transparent deployment of machine learning systems.

Ecological Interface Design (EID) stresses that interfaces should make the underlying invariants, constraints, and structure of a system perceptually available, so the relevant affordances of actions become cognitively accessible for a system operator (Vernon et al., 2002, p. 224). By visually signaling low-contrastivity predictions or classes, the Risk Mapper operationalizes this guidance—making otherwise hidden model uncertainties and reasoning limitations transparent for user review and intervention. As highlighted by Miller (2018), meaningful explanations in artificial intelligence must clarify why a particular outcome is chosen over alternatives, particularly under conditions of ambiguity or uncertainty. Integrating explanation contrastivity into system design helps satisfy these criteria and supports deeper user engagement with model strengths and weaknesses.

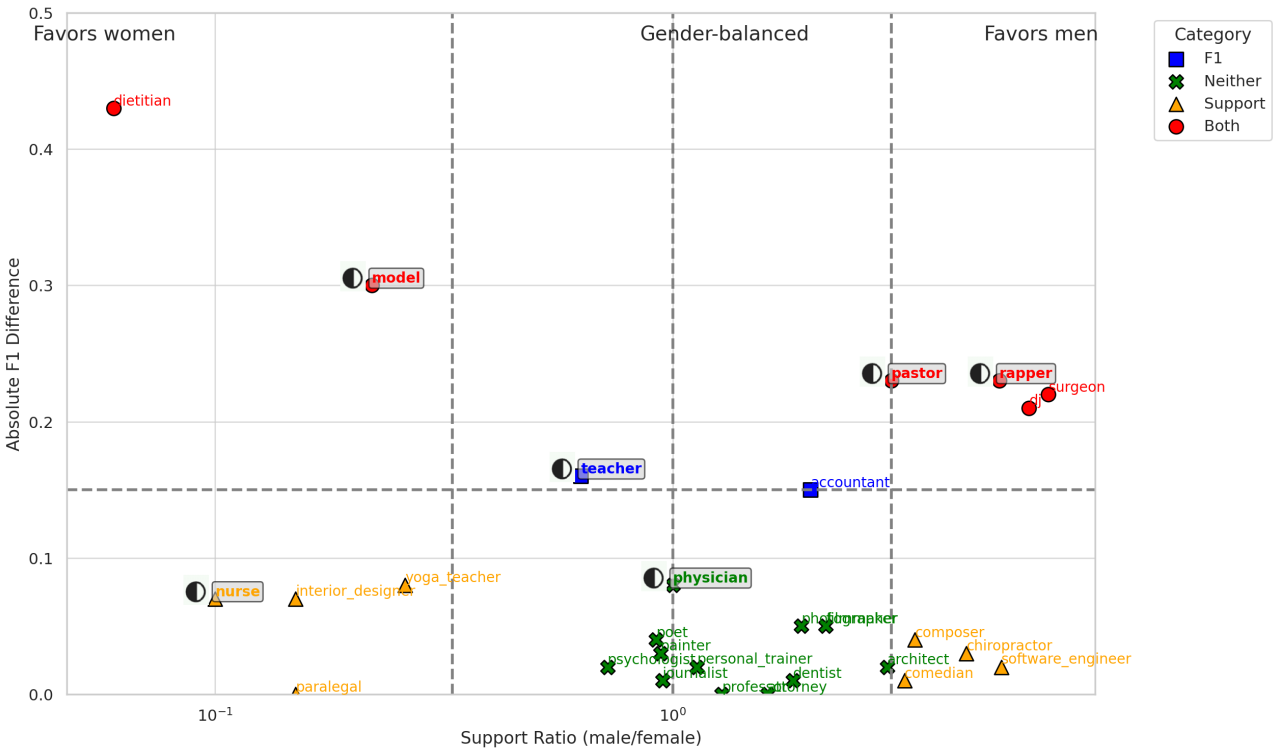

Figure 6.6: Occupational classes with statistically significant lack of contrastivity

In the diagram (see Figure 6.6), occupational classes with statistically significant lack of contrastivity are highlighted using a bold, high-contrast border. This explicit visual cue enables

immediate identification of classes that are not merely affected by data imbalance or performance disparity, but also exhibit ambiguous or weak model rationales—directing attention to areas that require deeper, instance-level investigation. As a result, the Risk Mapper operationalizes explainability as a critical risk factor, supporting the layered transparency and auditability principles advocated by EID.

6.4.3 Detecting hidden contrastivity risks in non-significant classes

While statistical tests, such as those producing p-values, are commonly used to flag professions with potential fairness issues, our analysis reveals that there exist classes where the risk may remain undetected by conventional significance thresholds. As evidenced in Table A.3 in Appendix A, several professions display a non-significant p-value, yet demonstrate notable disparities in F1 scores or pronounced support imbalances between groups. These cases are especially relevant when considering the broader remit of contrastivity and transparent model reasoning.

To address this complexity, we perform an extensive analysis of *borderline contrastivity risk*—identifying cases where the quantitative indicators suggest unequally reasoned outcomes, though these do not cross conventional significance thresholds. Such an approach is critical in high-stakes domains, as apparent fairness (per conventional cut-offs) may mask subtler, model-specific forms of risk or bias.

Borderline Class Identification. To systematically flag these borderline classes for further review, we introduce an algorithmic criterion based on three dimensions: statistical insignificance (p-value), relevant model performance difference (F1 difference), and group support ratio. A class is included in the borderline risk set if it satisfies the following:

Definitions. Let

- p = p-value for the statistical fairness test;
- D_{F_1} = absolute difference in F_1 between groups;
- R_s = support ratio (greater-to-lower group count);

- δ = threshold for a meaningful F_1 difference (set to 0.07 for this study).

Then, a class is flagged as *borderline risk* if:

$$p > 0.05 \ \wedge \ |D_{F_1}| \geq \delta \ \wedge \ \big((R_s < 3 \wedge R_s > 0.33) \ \vee \ R_s \geq 3 \ \vee \ R_s \leq 0.33\big).$$

This criterion leverages the performance-difference constant ($\delta = 0.07$) tailored to the accuracy and class-balance dynamics observed in this study's model. For other models or datasets, **the selection of δ must be empirically justified and adapted**, as model-specific factors—such as overall accuracy, class representation, and decision-threshold sensitivity—can markedly alter the boundary at which differences are meaningful or actionable.

Practical Implications

By applying this criterion, we ensure that classes with potentially meaningful disparities - borderline in terms of contrastivity and fairness - receive additional scrutiny, regardless of their statistical significance under traditional tests. This aligns with best practices from EID and cognitive systems engineering, emphasizing that system boundaries and uncertainties should be made observable and interpretable for oversight and risk management.

Classes that surpass these thresholds - but do not necessarily reach statistical significance (as reflected by their p-values) - are considered borderline contrastivity classes. These are depicted on the diagram, presented on Figure 6.7, as points lying near or beyond the dashed boundaries, often marked with specific colors or shapes corresponding to their category (e.g., blue squares for significant F1 difference, orange triangles for support imbalance, and red circles for both).

In summary, our approach operationalizes an additional diagnostic layer for model transparency: surfacing borderline classes that, while not statistically significant by conventional means, might still represent a reasoning risk. This not only supports accountable and fair decision support in risky domains but enables domain specialists to proactively track and audit subtle sources of potential unfairness.

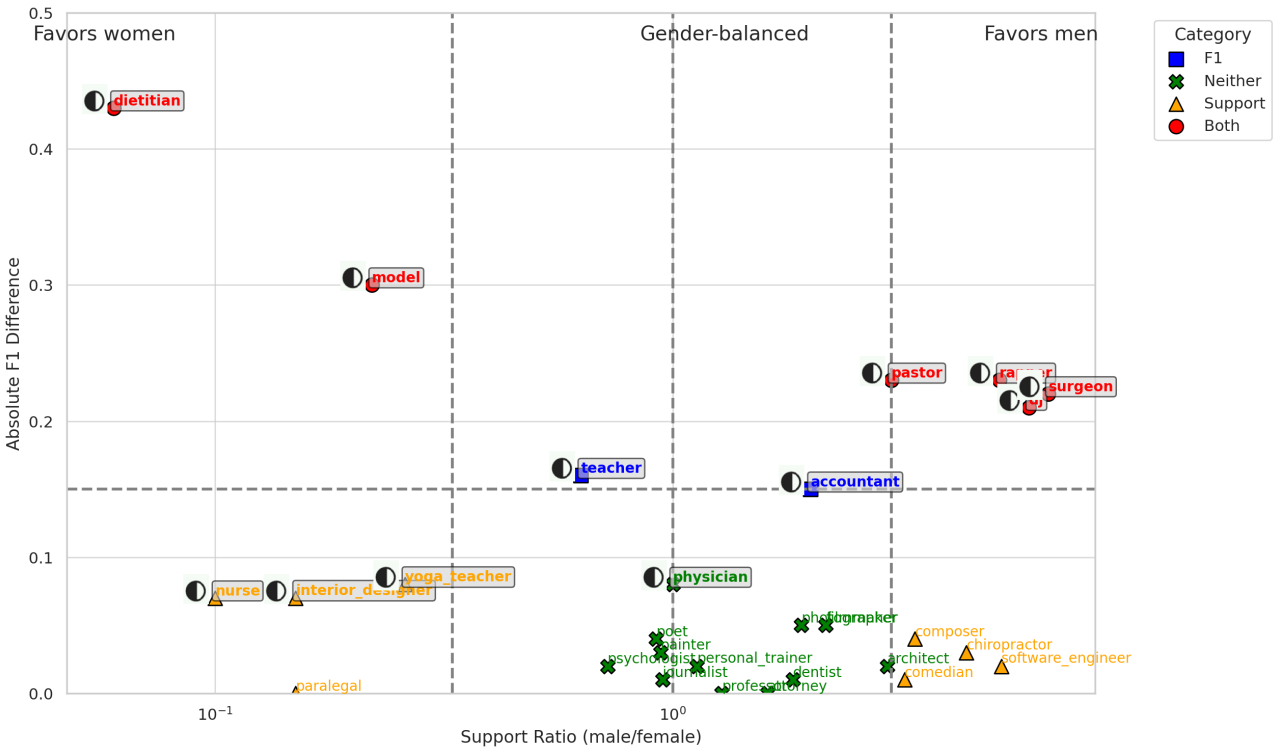

Figure 6.7: Occupational classes with lack of contrastivity including those where the statistical significance likely affected by data scarcity

6.5 Proximity Compatibility Principle–informed visualization for fairness monitoring and reasoning

In this section, we deepen the account of our interface design process and motivate a specific visualization technique that achieves efficiency by orchestrating multiple perceptual dimensions on a single display. Building on the Proximity Compatibility Principle (PCP) (Wickens & Andre, 1990), we show how color, spatial position, and size—augmented by shape and boundary cues—can be combined to support both integrated judgments and selective diagnostics. The goal is to let operators perceive overall fairness status immediately while being able to interrogate causes of inconsistencies without costly context switching.

We describe how spatial mappings encode the joint state of representation balance and performance parity; how color and shape function as separable channels for diagnostic categorization; how size modulates salience for workload balancing; and how subtle boundary lines externalize policy thresholds to reduce cognitive transformation. Together, these choices enable rapid pre-attentive appraisal, minimize visual search, and provide a scalable pathway from summary awareness to detailed reasoning within the same interactive frame.

6.5.1 Risk Mapper group fairness dashboard interface

The system dashboard for group-fairness assessment was designed in accordance with the PCP and related guidance from EID (Vicente, 1999). PCP posits that when a task requires the user to integrate multiple, tightly coupled variables, the associated information should be displayed with high spatial proximity and encoded to promote holistic perception; conversely, when the task demands selective attention, the display should employ separable codes that minimize interference and search (Ware, 2019; Wickens & Carswell, 1995; Wickens & Hollands, 2000).

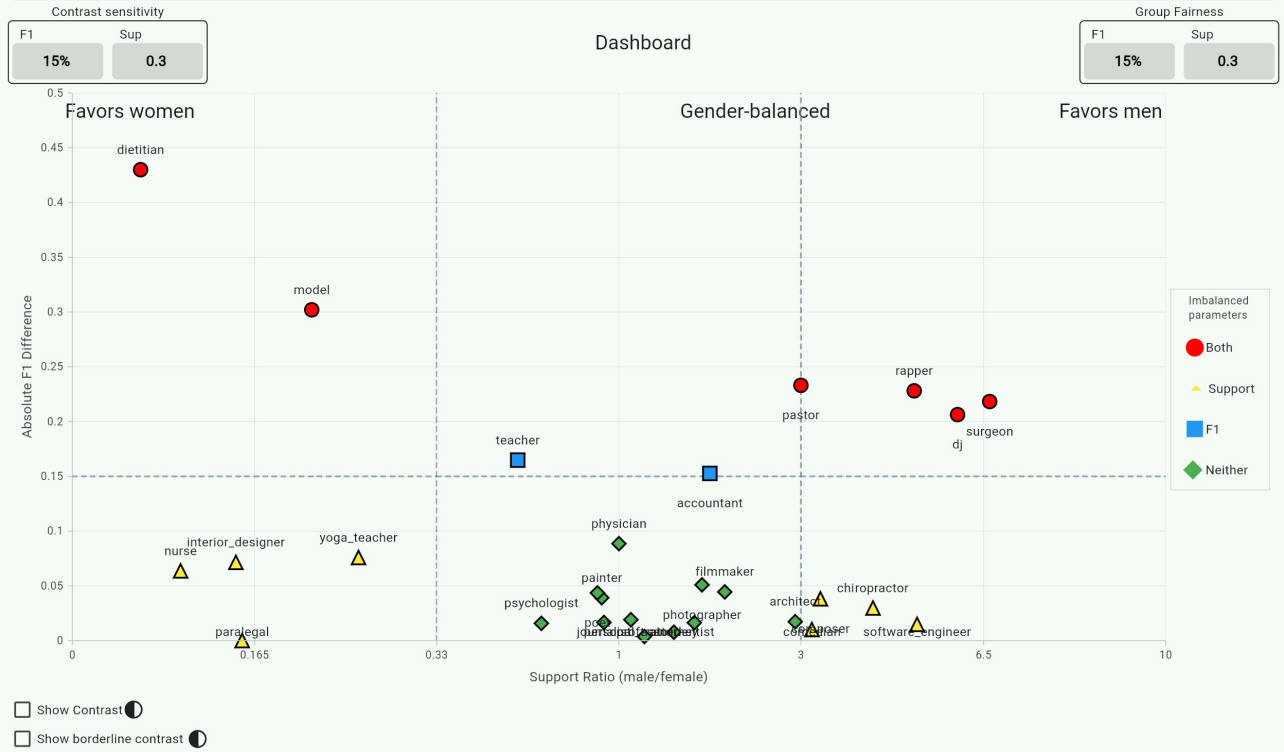

Figure 6.8: The spatial integration for balance representation

The primary operator judgment in this context involves a joint assessment of representation balance (support ratio, male/female) and performance parity (absolute ΔF_1) for each occupation (Figure 6.8). Because these variables are tightly coupled in the decision, the interface co-locates them within a single graphical space: the support ratio is mapped to the x-axis (centered at $x = 1$), and ΔF_1 is mapped to the y-axis (centered at $y = 0$). This spatial integration enables users to perceive the combined state directly from point position, reducing the need for mental combination and thereby aligning with PCP's recommendations for integrated tasks (Wickens & Hollands, 2000).

To externalize constraints and reduce cognitive transformation, policy thresholds are represented as dashed vertical and horizontal boundaries. These lines partition the display into perceptually contiguous regions (e.g., safe, caution, risky), facilitating rapid pre-attentive classification through emergent grouping (Bennett & Flach, 1992; Vicente, 1999). After the integrated read, users often need to diagnose the source of imbalance. To support this selective process without disrupting the integrated spatial coding, imbalance types are encoded using separable visual channels—hue and shape—to distinguish cases driven by "Support," "F1," "Both," or "Neither." Such separable coding mitigates interference with the spatial variables while enabling efficient drill-down (Ware, 2019; Wickens & Carswell, 1995).

Explanatory elements are kept in functional proximity to the data. The legend is placed within the plotting area and concise status panels appear along the top edge, minimizing eye travel and lookup costs and maintaining co-location of interpretive aids with judged objects (Wickens et al., 2004). Visual economy is preserved by labeling only salient points, thereby maintaining attentional proximity to high-value items and avoiding crowding that would diminish the benefits of proximity and impair pre-attentive processing (Ware, 2019).

The design further leverages emergent features—distance from $x = 1$ and $y = 0$, and clustering within threshold-defined zones—so that severity and category can be perceived rapidly and holistically, consistent with PCP's emphasis on perceptual integration (Bennett & Flach, 1992; Vicente, 1999). The reasoning exploration is presented as an on-demand option, to balance integration with flexibility (Figure 6.9). This keeps high proximity when needed while avoiding unnecessary code mixing during routine monitoring (Wickens & Hollands, 2000). Region labels such as "Favors women," "Gender-balanced," and "Favors men" are embedded along the spatial continuum to align spatial semantics with the decision frame, reducing interpretive recoding and reinforcing direct perception–action coupling (Bennett & Flach, 1992).

Finally, the interface was evaluated against PCP's core compliance criteria: tightly coupled variables are integrated spatially, while diagnostic attributes are expressed through separable channels; grouping, labeling, and placement were tuned to minimize interference and visual search. This proximity-interference balance operationalizes PCP within a practical fairness-monitoring display.

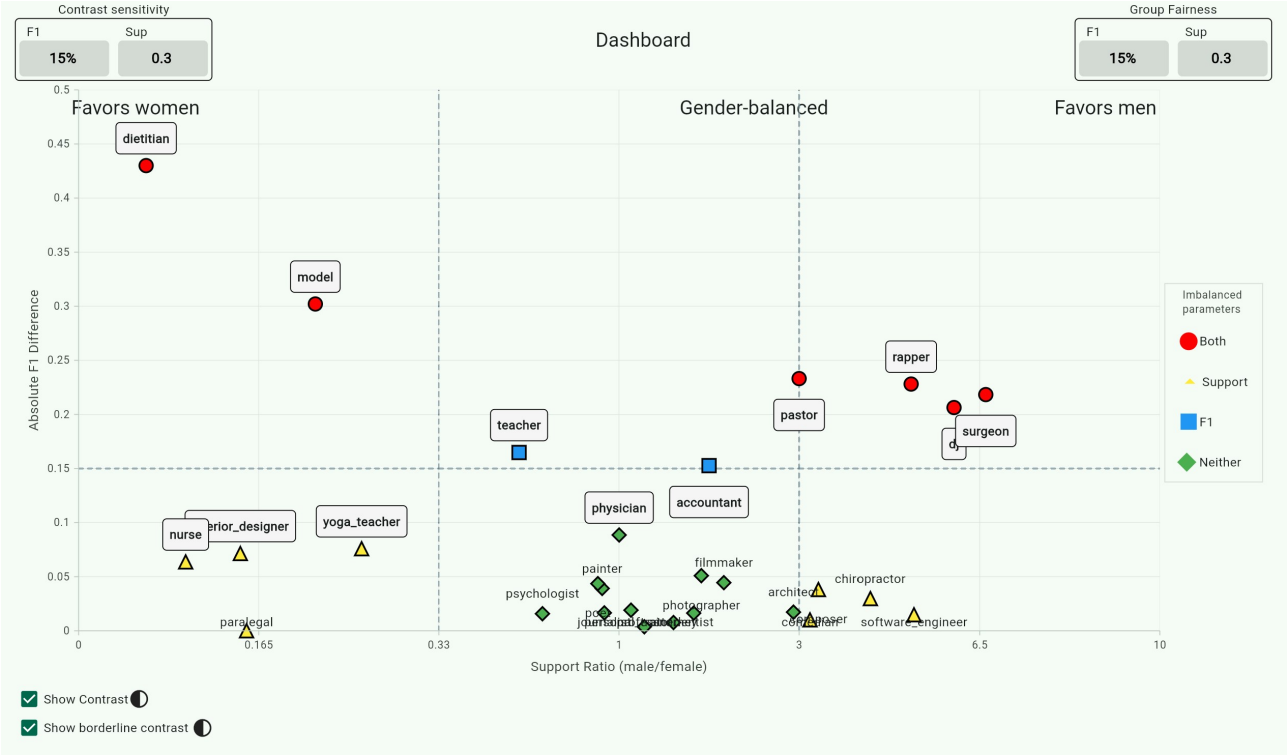

Figure 6.9: On-denamd reasoning representation for flexibility and increasing proximity

6.5.2 Individual prediction reasoning assessment interface

The assessment of an individual model prediction is important for understanding the reasoning and fairness associated with that specific outcome, in contrast to group assessment, which demonstrates characteristics based on the group to which the individual prediction belongs. For the assessment of individual predictions, we propose a Contrastivity of Explanation metric, described in detail in 4.2.3 of this research. This metric characterizes the reasoning behind individual predictions; however, differences between demographic groups reveal group disparities. Consequently, the measure exhibits a dual nature, which motivated its implementation in both variations. Contrastivity of Explanation, presented as the "Contrast" feature on the Risk Mapper dashboard, constitutes the implementation of the group-level characteristic of this metric.

Because each individual prediction is associated with textual features that influence the prediction, it is important to link the model's outcome (predicted occupation class) to the specific text on which this prediction was made. As the textual element can be quite long, it is important to plan the design of the individual-prediction assessment so that all necessary features are presented; however, the number of features should not be so large as to distract the

operator or overload them with information irrelevant to the specific context.

For planning the design of the individual prediction assessment system function, we employed PCP (Wickens & Carswell, 1995) approach that allows us to differentiate the beneficial type of display proximity either for mental integration of the information and for focused attention. In the paradigm of proximity compatibility, the assessment of group predictions is based on emergent features that combine two dimensions (e.g., disparity and performance) and therefore benefits from co-located, integrative displays; by contrast, the individual prediction is distinguished as a single object requiring focused attention to its explanatory evidence.

Thus, in our implementation we separate the portfolio-level Risk Mapper representation from the exemplar-level reasoning view (Figure 6.10): the former supports high-proximity integration among many items on a shared coordinate system, while the latter reduces interference and isolates the individual prediction with its local metrics. This separation is consistent with empirical findings that when tasks shift from multi-component comparison to single-object diagnosis, performance is improved by lowering physical proximity (Wickens & Carswell, 1995, p. 3).

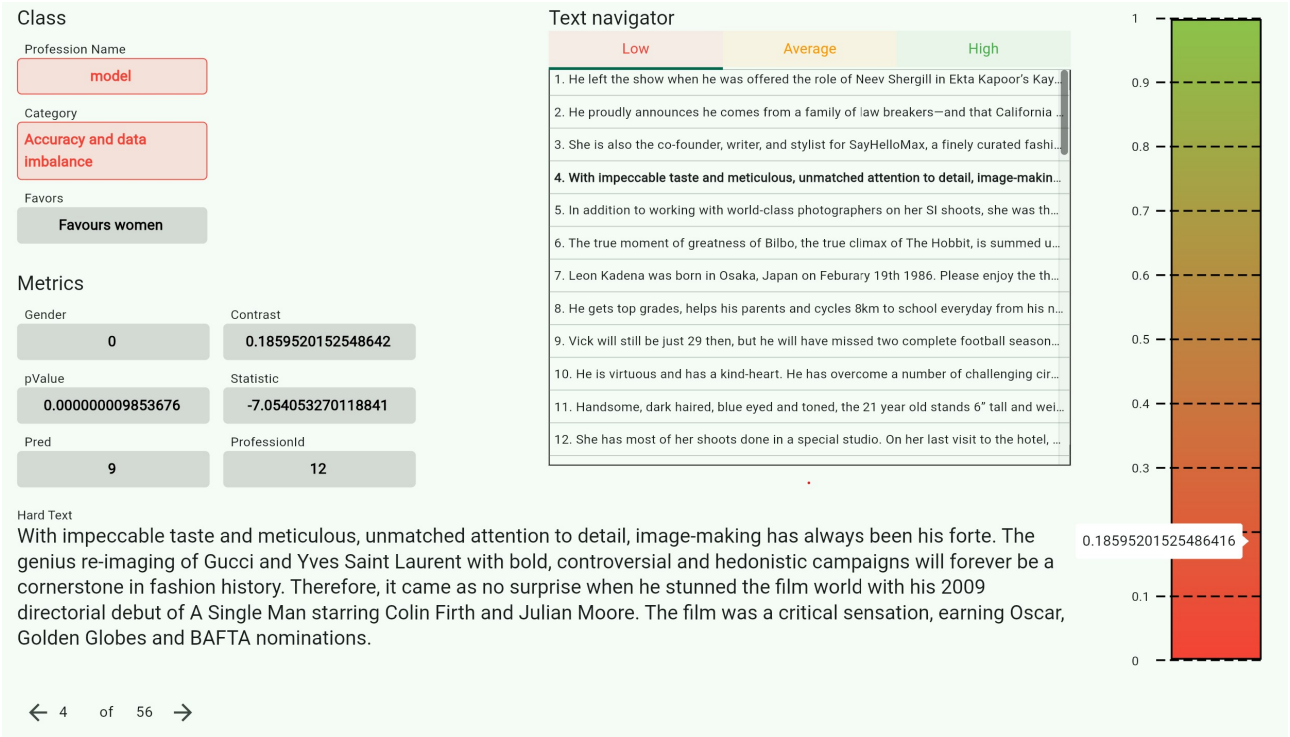

Figure 6.10: Individual prediction reasoning assessment display view

We also provide a continuous, vertically oriented *Contrastivity of Explanation* scale that maps the metric to a perceptually ordered red–amber–green ramp and a movable pointer; this preserves physical proximity between value, legend, and exemplar while keeping the color channel separable from other indicators. At the same time, we avoid forcing emergent, configural features in the individual view that would bias the display toward integration when the task is to inspect one prediction and its textual evidence. Where limited integration is beneficial (linking the predicted class to its main textual exemplar), we use objectness judiciously by grouping only spatially commensurate dimensions and keeping reference cues adjacent, while maintaining separate hues for distinct indicators. This matches established results: using the same color helps an operator combine information across many items, while using different colors helps the operator focus on one thing; even when data are shown as a single "object," giving its parts separate colors can improve focused attention with minimal cost to speed (Wickens & Andre, 1990; Wickens & Carswell, 1995).

A navigator pane is used to group the individual prediction by its contrast level, organizing the case's text snippets into Low (red), Average (amber), and High (green) bands based on their Contrastivity of Explanation. Each row in the pane corresponds to a sentence or fragment from the source text and is placed beneath the color-coded band that matches its contrastivity score, which is calibrated to the vertical contrastivity scale on the right. This design enables rapid, pre-attentive grouping through color proximity while keeping items separable for focused reading: analysts can scan the green band first to locate the strongest explanatory evidence, then review amber and red as needed. Selecting a row highlights the corresponding passage in the full text (and vice versa), maintaining virtual proximity between the navigator and context. In line with the PCP, the pane leverages separate colors to reduce search costs and code interference for single-case diagnosis, while preserving tight spatial proximity among related elements (labels, thresholds, and values) only where brief, local integration is required (Wickens & Carswell, 1995).

We align the second-level reasoning display with PCP and with visual momentum—the maintenance of spatial and semantic continuity across views to minimize reorientation costs (Wickens & Carswell, 1995; Wickens et al., 2004). Coupled constructs—evidence spans and

contrast magnitude—are co-located so severity can be read pre-attentively from a vertical gauge, while bolded text highlights serve as a separable diagnostic code. To sustain visual momentum across the overview and detail, we keep anchors stable: the same left–right semantics, the same occupation labels in the same horizontal order, and identical color semantics. Here we use a standard "RAG" risk scheme—Red = low reasoning quality/high risk, Amber = average, Green = sufficient—and we avoid introducing new hues; imbalance types ("Both/Support/F1") are conveyed via shape or icon overlays to preserve the RAG meaning. We add context overlap by pinning a small inset of the global risk map in the reasoning view that highlights the selected node; conversely, node tooltips in the overview include a mini RAG histogram of reasoning quality. Landmarks and breadcrumbs further support object constancy and path memory: the Class panel (profession name, category, favors) persists across views, helping users stay oriented as they move between integrated appraisal and focused diagnosis (Bennett & Flach, 2012; Vicente, 1999). Complete source code and configuration files are available at `github.com/XeniyaVl-git/AI_Transparency/tree/main/metrik_viewer_x` (release `v1.0.1`), archived at DOI: `10.5281/zenodo.16925254`.

To summarize, displays for individual prediction assessment—each associated with a single object—enhance operator performance by supporting focused attention. Moreover, the color coding that integrates contrastivity of explanation as a complementary, instance-aware signal extends the assessment beyond aggregate parity checks to the quality of the model's reasoning, revealing classes and cases in which rationales are weak or ambiguous even when representation or accuracy appears acceptable. Incorporating individual-level assessment thus advances the primary goal of the AI system and, together with the Risk Mapper, constitutes a virtual ecology for fairness evaluation, complemented by greater transparency of AI predictions (Bennett & Flach, 1992; Vernon et al., 2002; Vicente, 1999).

Chapter 7

Discussion and Conclusion

The proposed EID-based framework for AI model transparency extends existing explainability practices by embedding post-hoc interpretability techniques within a constraint-based design structure. Rather than treating post-hoc methods as isolated add-ons, the framework integrates them into an abstraction hierarchy that links lower-level model elements—such as features, coefficients, and attention weights—to higher-level system functions and constraints. This layered mapping enables end-users to navigate seamlessly between granular algorithmic details and abstract, goal-oriented representations, thereby fostering a more comprehensive understanding of model behaviour.

This integration addresses a key limitation identified in the literature: most transparency work emphasises algorithmic outputs or local explanations in isolation (Ribeiro et al., 2016; Zeng et al., 2017) without embedding them into a coherent system representation. By contrast, the proposed framework positions post-hoc outputs within the broader operational and societal context, making fairness-relevant factors, causal dependencies, and procedural reasoning directly observable in the interface. In doing so, it responds to calls for AI governance approaches that incorporate psychological, societal, and philosophical perspectives (Lopes, 2025; Nakao et al., 2023; Yang et al., 2025), ensuring that transparency mechanisms account for both technical validity and human perception.

Leveraging EID principles allows the system to remain adaptable to dynamic, uncertain environments while maintaining interpretability across abstraction levels. This adaptability,

rooted in constraint-based design, aligns with theories of distributed cognition and joint cognitive systems by supporting fluid human–AI collaboration and enabling users to adjust their strategies in response to changing operational constraints (Naikar et al., 2023).

Nonetheless, the conceptual framework, while demonstrating reliance on abstraction hierarchy modelling and constraint visualisation, may require extension to incorporate a broader range of mathematical and statistical approaches for ensuring model fairness and explainability. Validation of this concept across diverse domains, using qualitative methods, may provide insights into its applicability in different implementations and system roles, and will be essential for establishing the framework's practical benefits for trust, accountability, and equitable decision-making.

In summary, this work demonstrates that post-hoc explainability methods, when systematically integrated into a constraint-based EID architecture, can move beyond isolated transparency enhancements to form part of a unified, multi-level interpretability framework. By linking the detailed mechanics of model predictions to overarching system purposes and constraints, the framework advances both operational transparency and the societal trustworthiness of AI systems.

In conclusion, this thesis advances AI transparency from principle to practice by developing an EID-grounded framework and validating it across the full AI-modeling pipeline. We began by articulating the core transparency objectives and EID principles, clarifying how explainability, interpretability, and disclosure relate to stakeholder needs, governance requirements, and fairness concerns. Building on this foundation, we detailed methodological components—data curation, labeling assumptions, and model selection—emphasizing how these upstream choices shape the scope of explanations and the validity of downstream evaluations. We then operationalized the framework with a domain application: training a BERT-based, black-box classifier to predict occupations from short textual samples such as bios and job descriptions.

Treating the model as opaque, we introduced explanation tools centered on an *Explanation Contrastivity* metric designed to compare model rationales across exemplars, detect inconsistencies,

and align explanatory evidence with task-relevant signals (see Miller, 2018). Empirically, the metric distinguished fair from unfair predictions, revealed set-level imbalances, and surfaced statistical disparities, correctly identifying 66.7% of class exemplars with accuracy-related bias; its sensitivity is expected to strengthen with larger exemplar counts, particularly under class imbalance.

We complemented this with a fairness analysis that situates explanation quality within broader equity considerations, connecting observed disparities to actionable diagnostic signals rather than isolated performance statistics. Crucially, the proposed EID-based framework extends existing explainability practices by embedding post-hoc interpretability techniques within a constraint-based design structure: instead of treating explanations as isolated add-ons, it integrates them into an abstraction hierarchy that links lower-level model elements (features, coefficients, attention weights) to higher-level system functions and constraints, enabling fluid navigation between granular algorithmic details and goal-oriented representations. This integration addresses a key limitation in the literature—where algorithmic outputs or local explanations are often presented in isolation—by positioning post-hoc outputs within broader operational and societal contexts so that fairness-relevant factors, causal dependencies, and procedural reasoning become directly observable, responding to calls for governance (Lopes, 2025; Nakao et al., 2023; Ribeiro et al., 2016; Yang et al., 2025; Zeng et al., 2017).

Utilizing EID principles within a constraint-based architecture preserves adaptability in dynamic, uncertain environments while maintaining interpretability across abstraction levels, aligning with theories of distributed cognition and joint cognitive systems to support fluid human–AI collaboration and strategy adjustment under changing constraints (Naikar et al., 2023). While the concept demonstrates the value of abstraction hierarchy modeling and constraint visualization, it may benefit from broader mathematical and statistical extensions for fairness and explainability, and from qualitative validation across diverse domains to establish practical benefits for trust, accountability, and equitable decision-making.

Collectively, these results establish a robust, generalizable path for transparent and trustworthy AI—one that integrates explanation contrastivity, fairness evaluation, and domain-grounded validation.

7.1 Limitations and Future Work

This work demonstrates the strength of a cognitive–engineering (CWA/EID) approach for creating transparent AI systems by explicitly linking low–level model signals to higher–level goals, constraints, and affordances. The framework integrates post–hoc explanations into an abstraction hierarchy and renders them perceptible through constraint–based visualization, providing operators with actionable evidence for assessment and oversight. However, validating this approach across additional domains, data regimes, and model families is necessary to establish its generality and boundary conditions.

A key practical limitation concerns data sparsity. Although the proposed *Explanation Contrastivity* construct is designed to be broadly applicable to classification tasks, its sensitivity can be attenuated when minority classes or demographic subgroups contain few exemplars. Small sample sizes weaken statistical power and may obscure true differences in contrastivity, while also making estimates unstable. Future work should examine settings with larger and more balanced datasets, where contrastive reasoning signals can be quantified with higher reliability and calibrated against stronger ground truth.

The current instantiation does not probe deeper linguistic structure. Many biases in text classification arise from non–trivial linguistic phenomena (e.g., discourse, pragmatics, syntax, sociolinguistic markers) that are not captured by surface–level token attributions alone. Extending the approach to incorporate richer linguistic analyses (e.g., dependency and semantic role structures, discourse cues, counterfactual edits) would help validate the metric and strengthen sense–making for textual analysis and classification.

Additional limitations include: the focus on a single task/domain and primarily transformer–based models; heuristic choices for visualization thresholds that merit sensitivity analysis; associative (not causal) explanations that may conflate correlation with mechanism; potential dependence of contrastivity on feature correlations and preprocessing choices. Addressing these issues—through cross–domain replication, causal and robustness analyses, user experiments, richer linguistic instrumentation, and sustained lifecycle monitoring—constitutes an important agenda for future research.

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 Inf Technol*, *26*(2), 31. https://doi.org/10.1007/s10676-024-09746-w

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. Retrieved August 21, 2025, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/

Barsever, D., Singh, S., & Neftci, E. (2020). Building a Better Lie Detector with BERT: The Difference Between Truth and Lies. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206937

Bennett, K. B., & Flach, J. M. (1992). Graphical Displays: Implications for Divided Attention, Focused Attention, and Problem Solving. *Hum Factors*, *34*(5), 513–533. https://doi.org/10.1177/001872089203400502

Bennett, K. B., & Flach, J. M. (2012). Visual momentum redux. *International Journal of Human-Computer Studies*, *70*(6), 399–414. https://doi.org/10.1016/j.ijhcs.2012.01.003

Burns, C., Bryant, D., & Chalmers, B. (2000). A work domain model to support shipboard command and control. *SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions' (Cat. No.00CH37166)*, *3*, 2228–2233. https://doi.org/10.1109/ICSMC.2000.886447

Chouldechova, A. (2016, October). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments [arXiv:1610.07524 [stat]]. https://doi.org/10.48550/arXiv.1610.07524

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting [arXiv:1901.09451 [cs]]. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128. https://doi.org/10.1145/3287560.3287572

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]]. https://doi.org/10.48550/arXiv.1810.04805

Doshi-Velez, F., & Kim, B. (2017, March). Towards A Rigorous Science of Interpretable Machine Learning [arXiv:1702.08608 [stat]]. https://doi.org/10.48550/arXiv.1702.08608

Erdoğanyılmaz, C., Mengünoğul, B., & Balci, M. (2023). Unveiling the Black Box: Investigating the Interplay between AI Technologies, Explainability, and Legal Implications. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 569–574. https://doi.org/10.1109/UBMK59864.2023.10286653

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition* (1st). Psychology Press. https://doi.org/10.4324/9781315740218

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hepenstal, S. (2023). *Designing a Transparent Conversational System for Intelligence Analysis.pdf* [Doctoral dissertation]. https://openrepository.aut.ac.nz/items/8341eece-13bd-4081-9fc4-4b83fefd6150

Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., & Wong, B. L. W. (2019). Algorithmic Transparency of Conversational Agents. *Los Angeles*.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

Hollnagel, E. (2008). Joint cognitive systems: Patterns in cognitive systems engineering. *Ergonomics*. https://doi.org/10.1080/00140130701223774

Lemieux, V., Rahmdel, P. S., Walker, R., Wong, B. L. W., & Flood, M. (2014). Clustering Techniques And their Effect on Portfolio Formation and Risk Analysis. *Proceedings of the International Workshop on Data Science for Macro-Modeling*, 1–6. https://doi.org/10.1145/2630729.2630749

Lopes, G. P. (2025). Bias in Adjudication and the Promise of AI: Challenges to Procedural Fairness. *Law Tech Hum*, *7*(1), 47–67. https://doi.org/10.5204/lthj.3812

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–158. https://doi.org/10.1145/2339530.2339556

Lundberg, S., & Lee, S.-I. (2017, November). A Unified Approach to Interpreting Model Predictions [arXiv:1705.07874 [cs]]. Retrieved October 20, 2024, from http://arxiv.org/abs/1705.07874

Miller, T. (2018, August). Explanation in Artificial Intelligence: Insights from the Social Sciences [arXiv:1706.07269 [cs]]. https://doi.org/10.48550/arXiv.1706.07269

Moll, I. (2022). The Fourth Industrial Revolution: A New Ideology. *tripleC*, *20*(1), 45–61. https://doi.org/10.31269/triplec.v20i1.1297

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Morzhov, S. (2020). Avoiding Unintended Bias in Toxicity Classification with Neural Networks. *2020 26th Conference of Open Innovations Association (FRUCT)*, 314–320. https://doi.org/10.23919/FRUCT48808.2020.9087368

Naikar, N. (2016, April). *Work Domain Analysis: Concepts, Guidelines, and Cases* (0th ed.). CRC Press. https://doi.org/10.1201/b14774

Naikar, N. (2017). Cognitive work analysis: An influential legacy extending beyond human factors and engineering. *Applied Ergonomics*, *59*, 528–540. https://doi.org/10.1016/j.apergo.2016.06.001

Naikar, N., Brady, A., Moy, G., & Kwok, H.-W. (2023). Designing human-AI systems for complex settings: Ideas from distributed, joint, and self-organising perspectives of sociotechnical systems and cognitive work analysis. *Ergonomics*, *66*(11), 1669–1694. https://doi.org/10.1080/00140139.2023.2281898

Nakao, Y., Strappelli, L., Stumpf, S., Naseer, A., Regoli, D., & Gamba, G. D. (2023). Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. *International Journal of Human–Computer Interaction*, *39*(9), 1762–1788. https://doi.org/10.1080/10447318.2022.2067936

Narayanan, D., Nagpal, M., McGuire, J., Schweitzer, S., & De Cremer, D. (2024). Fairness Perceptions of Artificial Intelligence: A Review and Path Forward. *International Journal of Human–Computer Interaction*, *40*(1), 4–23. https://doi.org/10.1080/10447318.2023.2210890

Norman, D. (1986, January). Cognitive Engineering. In *User Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 31–61). https://doi.org/10.1201/b15703-3

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. (3). https://doi.org/10.1109/TSMC.1983.6313160

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. *IEEE Trans. Syst., Man, Cybern.*, *SMC-15*(2), 234–243. https://doi.org/10.1109/TSMC.1985.6313353

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. Wiley.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why Should I Trust You?": Explaining the Predictions of Any Classifier [arXiv:1602.04938 [cs]]. https://doi.org/10.48550/arXiv.1602.04938

Schwab, K. (2016). The Fourth Industrial Revolution. https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

Tieu, A., & Naikar, N. (2022). Visualizations for human-machine teams in complex environments: Design concepts and review of current approaches. *2022 IEEE 3rd International Conference*

on Human-Machine Systems (ICHMS), 1–7. https://doi.org/10.1109/ICHMS56717.2022.9980811

U.S. Government Accountability Office. (2021, June). Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities — U.S. GAO. Retrieved July 9, 2024, from https://www.gao.gov/products/gao-21-519sp

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August). Attention Is All You Need [arXiv:1706.03762 [cs]]. https://doi.org/10.48550/arXiv.1706.03762

Vernon, D., Reising, C., & Sanderson, P. M. (2002). Ecological Interface Design for Pasteurizer II: A Process Description of Semantic Mapping. *Hum Factors*, *44*(2), 222–247. https://doi.org/10.1518/0018720024497952

Vicente, K. J. (1995). Ecological Interface Design: A Research Overview. *IFAC Proceedings Volumes*, *28*(15), 623–628. https://doi.org/10.1016/S1474-6670(17)45302-X

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates.

Vicente, K. J., & Rasmussen, J. (1990). The Ecology of Human-Machine Systems II: Mediating 'Direct Perception' in Complex Work Domains. *Ecological Psychology*, *2*(3), 207–249. https://doi.org/10.1207/s15326969eco0203_2

Volkov, E. N., & Averkin, A. N. (2024). Local Explanations for Large Language Models: A Brief Review of Methods. *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, 189–192. https://doi.org/10.1109/SCM62608.2024.10554222

Ware, C. (2019). *Information visualization: Perception for design*. Morgan Kaufmann. https://doi.org/10.1016/C2009-0-62432-6

Wickens, C. D., & Andre, A. D. (1990). Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration. *Hum Factors*, *32*(1), 61–77. https://doi.org/10.1177/001872089003200105

Wickens, C. D., & Carswell, C. M. (1995). The Proximity Compatibility Principle: Its Psychological Foundation and Relevance to Display Design. *Hum Factors*, *37*(3), 473–494. https://doi.org/10.1518/001872095779049408

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Prentice Hall. https://books.google.co.nz/books?id=zvpiQgAACAAJ

Wickens, C. D., Lee, J., Gordon, S. E., & Liu, Y. D. (2004). *An introduction to human factors engineering*. Pearson Prentice Hall. https://books.google.co.nz/books?id=nGUItAEACAAJ

Wong, B. W., & Gulden, J. (2017). Risk Map as a Library Management Information Dashboard: A Case Study in Adapting a Configural Display. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 287–291. https://doi.org/10.1177/1541931213601553

Woods, D. (2016, January). Origins of Cognitive Systems Engineering. https://doi.org/10.1201/9781315572529-3

Yang, S. C.-H., Folke, T., & Shafto, P. (2025). The Inner Loop of Collective Human–Machine Intelligence. *Topics in Cognitive Science*, *17*(2), 248–267. https://doi.org/10.1111/tops.12642

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable Classification Models for Recidivism Prediction [arXiv:1503.07810 [stat]]. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(3), 689–722. https://doi.org/10.1111/rssa.12227

Zhao, Y., Wang, Y., & Derr, T. (2022, December). Fairness and Explainability: Bridging the Gap Towards Fair Model Explanations [arXiv:2212.03840 [cs]]. https://doi.org/10.48550/arXiv.2212.03840

Zini, J. E., & Awad, M. (2023). On the Explainability of Natural Language Processing Deep Models [Publisher: Association for Computing Machinery (ACM)]. *ACM Computing Surveys*, *55*, 1–31. https://doi.org/10.1145/3529755

Appendix A

Additional Results

Class	Precision	Recall	F1-score	Support
0	0.77	0.70	0.74	132
1	0.59	0.65	0.62	252
2	0.84	0.88	0.86	805
3	0.94	0.55	0.69	86
4	0.84	0.72	0.78	80
5	0.72	0.89	0.79	140
6	0.85	0.96	0.90	382
7	0.79	0.87	0.83	103
8	0.77	0.66	0.71	35
9	0.85	0.80	0.82	212
10	0.83	0.16	0.27	31
11	0.79	0.72	0.75	500
12	0.79	0.81	0.80	202
13	0.80	0.82	0.81	459
14	0.83	0.74	0.78	186
15	0.00	0.00	0.00	45
16	0.61	0.66	0.63	76
17	0.68	0.78	0.72	32
18	0.81	0.92	0.86	577
19	0.82	0.85	0.84	1063
20	0.67	0.80	0.73	171
21	0.87	0.87	0.87	2927
22	0.79	0.71	0.75	482
23	0.87	0.73	0.79	37
24	0.66	0.69	0.67	168
25	0.74	0.73	0.73	358
26	0.70	0.58	0.63	425
27	0.62	0.74	0.68	34
accuracy			0.81	10000
macro avg	0.74	0.71	0.72	10000
weighted avg	0.81	0.81	0.81	10000

Table A.1: Classification report for the model's prediction performance on the test set

Example text	**Predicted label**
He specializes in development economics, household economics, and personnel economics. In 2003 he received his Ph.D. in Economics from the London School of Economics. Previously, he has worked as an Assistant Professor of Economics at the University of Chicago, Graduate School of Business and also, as a consultant for the World Bank. Centre for Research and Analysis of Migration	**professor** (21)

Top 10 SHAP values			
Positive		**Negative**	
Value	**Word**	**Value**	**Word**
0.131608118	Professor	-0.00268	and
0.054959735	economics		
0.050864078	economics		
0.045798459	Research		
0.029489516	received		
0.019623723	.		
0.017984577	,		
0.017602113	of		
0.017000999	Chicago		
0.017000999	2003		

Table A.2: Example instance, model prediction, and top-10 SHAP contributions.

Null hypothesis	Profession	Significance (p)	*t*-Statistic	Decision
The mean value of contrastivity for both genders is similar	accountant	0.115	1.596387159	Fail to reject the null-hypothesis
	architect	0.153	1.440520827	Fail to reject the null-hypothesis
	attorney	0.990	0.012030177	Fail to reject the null-hypothesis
	chiropractor	0.723	0.359413627	Fail to reject the null-hypothesis
	comedian	0.525	0.644305351	Fail to reject the null-hypothesis
	composer	0.636	0.475691415	Fail to reject the null-hypothesis
	dentist	0.958	0.053121047	Fail to reject the null-hypothesis
	dietitian	0.130	1.794047733	Fail to reject the null-hypothesis
	dj	0.396	0.932647405	Fail to reject the null-hypothesis
	filmmaker	0.484	0.702243065	Fail to reject the null-hypothesis
	interior_designer	0.481	0.757669053	Fail to reject the null-hypothesis
	journalist	0.590	-0.538822360	Fail to reject the null-hypothesis
	model	9.9×10^{-10}	-7.05405327	**Reject the null-hypothesis**
	nurse	0.044	2.071597661	**Reject the null-hypothesis**
	painter	0.174	1.365694209	Fail to reject the null-hypothesis
	paralegal	0.384	0.933765343	Fail to reject the null-hypothesis
	pastor	0.043	2.107072776	**Reject the null-hypothesis**
	personal_trainer	0.509	0.667824199	Fail to reject the null-hypothesis
	photographer	0.889	0.140079008	Fail to reject the null-hypothesis
	physician	0.001	3.199064624	**Reject the null-hypothesis**
	poet	0.356	0.926274126	Fail to reject the null-hypothesis
	professor	0.474	0.716785080	Fail to reject the null-hypothesis
	psychologist	0.323	0.990151691	Fail to reject the null-hypothesis
	rapper	0.004	3.265096835	**Reject the null-hypothesis**
	software_engineer	0.185	1.352138753	Fail to reject the null-hypothesis
	surgeon	0.085	1.748097027	**Reject the null-hypothesis**
	teacher	0.000216	3.738508314	**Reject the null-hypothesis**
	yoga_teacher	0.713	0.378296693	Fail to reject the null-hypothesis

Table A.3: Gender comparison of explanation contrastivity by profession (Student's t-test).

Profession	Class	Precision	Recall	F1-score	Support
accountant	0	0.60	0.68	0.64	44
architect	1	0.54	0.70	0.61	64
attorney	2	0.82	0.90	0.86	309
chiropractor	3	1.00	0.50	0.67	16
comedian	4	0.93	0.68	0.79	19
composer	5	0.68	0.84	0.75	32
dentist	6	0.92	0.95	0.93	135
dietitian	7	0.84	0.88	0.86	97
dj	8	1.00	0.40	0.57	5
filmmaker	9	0.91	0.78	0.84	67
interior_designer	10	1.00	0.04	0.07	27
journalist	11	0.80	0.75	0.77	257
model	12	0.80	0.87	0.83	166
nurse	13	0.79	0.87	0.83	418
painter	14	0.92	0.69	0.79	96
paralegal	15	0.00	0.00	0.00	39
pastor	16	0.50	0.42	0.46	19
personal_trainer	17	0.73	0.73	0.73	15
photographer	18	0.76	0.90	0.83	198
physician	19	0.91	0.86	0.88	531
poet	20	0.79	0.70	0.74	89
professor	21	0.86	0.90	0.88	1283
psychologist	22	0.81	0.69	0.74	281
rapper	23	1.00	1.00	1.00	6
software_engineer	24	0.69	0.67	0.68	27
surgeon	25	0.55	0.55	0.55	47
teacher	26	0.70	0.68	0.69	261
yoga_teacher	27	0.68	0.78	0.72	27
accuracy				**0.82**	**4575**
macro avg		**0.77**	**0.69**	**0.70**	**4575**
weighted avg		**0.82**	**0.82**	**0.81**	**4575**

Table A.4: BERT model's classification report for female individuals

Profession	Class	Precision	Recall	F1-score	Support
accountant	0	0.79	0.80	0.79	88
architect	1	0.66	0.59	0.63	188
attorney	2	0.83	0.90	0.86	496
chiropractor	3	0.93	0.56	0.70	70
comedian	4	0.89	0.69	0.78	61
composer	5	0.73	0.86	0.79	108
dentist	6	0.89	0.95	0.92	247
dietitian	7	0.38	0.50	0.43	6
dj	8	0.88	0.70	0.78	30
filmmaker	9	0.83	0.76	0.79	145
interior_designer	10	0.00	0.00	0.00	4
journalist	11	0.78	0.74	0.76	243
model	12	0.61	0.47	0.53	36
nurse	13	0.96	0.63	0.76	41
painter	14	0.84	0.81	0.82	90
paralegal	15	0.00	0.00	0.00	6
pastor	16	0.70	0.68	0.69	57
personal_trainer	17	0.91	0.59	0.71	17
photographer	18	0.85	0.90	0.88	379
physician	19	0.77	0.82	0.80	532
poet	20	0.62	0.79	0.70	82
professor	21	0.87	0.90	0.88	1644
psychologist	22	0.82	0.70	0.76	201
rapper	23	0.85	0.71	0.77	31
software_engineer	24	0.64	0.67	0.66	141
surgeon	25	0.80	0.75	0.77	311
teacher	26	0.59	0.48	0.53	164
yoga_teacher	27	0.75	0.86	0.80	7
accuracy				**0.81**	**5425**
macro avg		0.72	0.67	0.69	5425
weighted avg		0.81	0.81	0.81	5425

Table A.5: BERT model's classification report for male individuals (test set)

Class	Profession	Support (female)	Support (male)	F1 (female)	F1 (male)	Larger set
3	chiropractor	16	70	0.67	0.70	male
4	comedian	19	61	0.79	0.78	male
5	composer	32	108	0.75	0.79	male
7	dietitian	97	6	0.86	0.43	female
8	dj	5	30	0.78	0.21	male
10	interior_designer	27	4	0.07	0.00	female
12	model	166	36	0.53	0.30	female
13	nurse	418	41	0.83	0.76	female
15	paralegal	39	6	0.00	0.00	female
16	pastor	19	57	0.69	0.23	male
23	rapper	6	31	0.77	0.23	male
24	software_engineer	27	141	0.68	0.66	male
25	surgeon	47	311	0.77	0.22	male
27	yoga_teacher	27	7	0.80	0.72	female

Table A.6: Professions with the highest gender imbalance in the training set. Shaded rows indicate occupations where one gender shows notably lower F1.

Class	Profession	F1 (female)	F1 (male)	Diff
0	accountant	0.64	0.79	0.15
7	dietitian	0.86	0.43	0.43
8	dj	0.57	0.78	0.21
12	model	0.83	0.53	0.30
16	pastor	0.46	0.69	0.23
23	rapper	1.00	0.77	0.23
25	surgeon	0.55	0.77	0.22
26	teacher	0.69	0.53	0.16

Table A.7: The accuracy difference for predictions made for two gender groups

Profession	F1 diff (%)	Support (M)	Support (F)	Support ratio	Meets F1	Ratio diff	Both	p value
accountant	15	88	44	2.00	✓			0.114647139
architect	2	188	64	2.94				0.152833056
attorney	0	496	309	1.61				0.990405281
chiropractor	3	70	16	4.38		✓		0.722828663
comedian	1	61	19	3.21		✓		0.524705065
composer	4	108	32	3.38		✓		0.636154740
dentist	1	247	135	1.83				0.957674177
dietitian	43	6	97	16.17	✓	✓	✓	0.130077318
dj	21	30	5	6.00	✓	✓	✓	0.395615635
filmmaker	5	145	67	2.16				0.488398362
interior_designer	7	4	27	6.75		✓		0.480874108
journalist	1	243	257	1.06				0.590205639
model	30	36	166	4.61	✓	✓	✓	0.0000000099
nurse	7	41	418	10.20		✓		0.044169779
painter	3	90	96	1.07				0.173709200
paralegal	0	6	39	6.50		✓		0.383812390
pastor	23	57	19	3.00	✓	✓	✓	0.042872476
personal_trainer	2	17	15	1.13				0.509407537
photographer	5	379	198	1.91				0.888666736
physician	8	532	531	1.00				0.001419673
poet	4	82	89	1.09				0.355624698
professor	0	1644	1283	1.28				0.473568919
psychologist	2	201	281	1.40				0.322656760
rapper	23	31	6	5.17	✓	✓	✓	0.004271675
software_engineer	2	141	27	5.22		✓		0.184960946
surgeon	22	311	47	6.62	✓	✓	✓	0.085486173
teacher	16	164	261	1.59	✓			0.000021617
yoga_teacher	8	7	27	3.86		✓		0.713327164

Table A.8: Summary of explanation contrastivity associated with model imbalances. Row colors follow the legend: LightGray (t-test significant, $p < 0.05$), SoftPink (borderline $0.05 \leq p < 0.10$), LegendBlue (outlier). Professions in red text indicate very small support for one gender (test result likely affected by data scarcity).

Appendix B

Additional Results

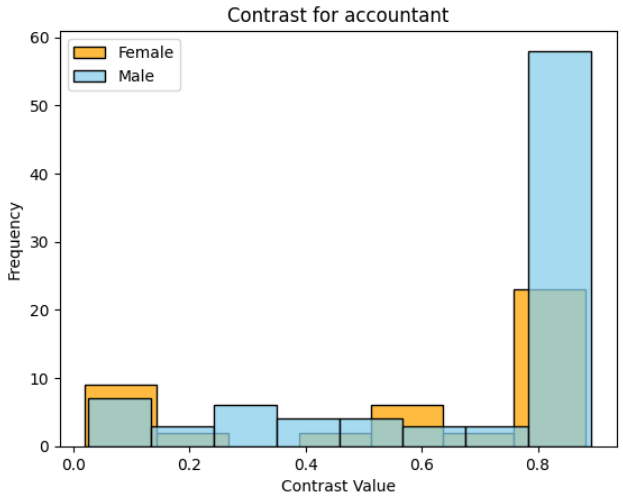

Figure B.1: The distribution of explanation contrastivity for an accountant.

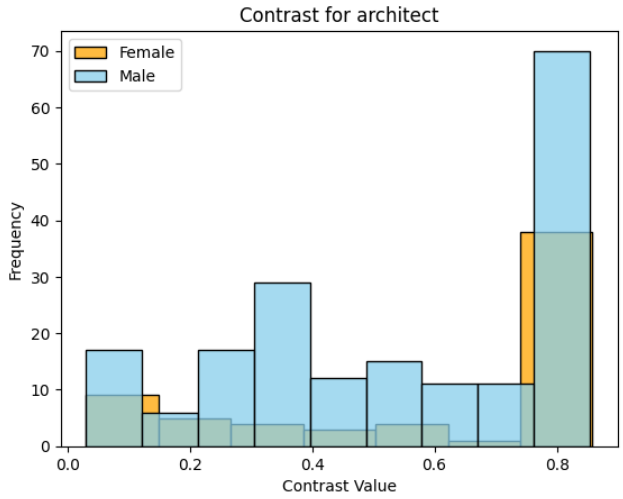

Figure B.2: The distribution of explanation contrastivity for an architect.

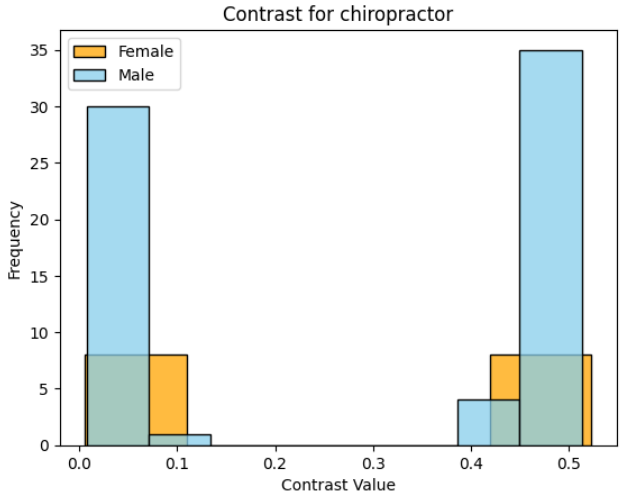

Figure B.3: The distribution of explanation contrastivity for a chiropractor.

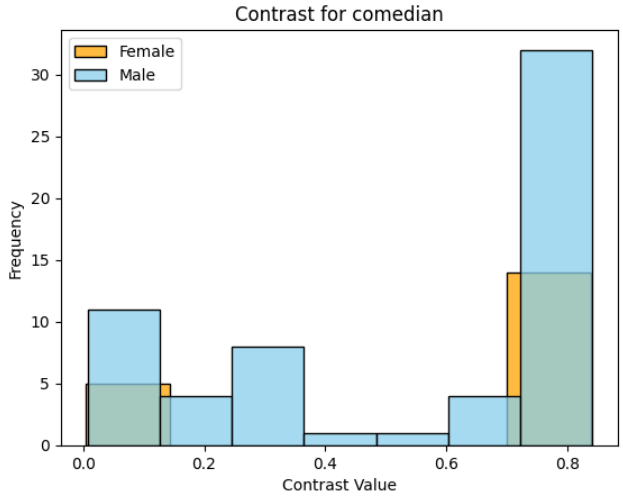

Figure B.4: The distribution of explanation contrastivity for a comedian.

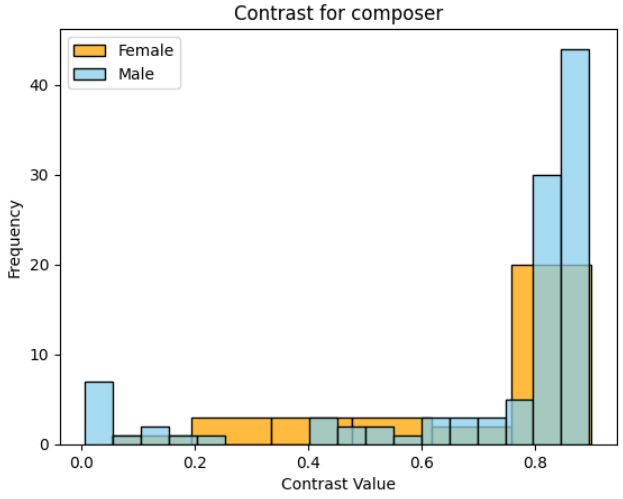

Figure B.5: The distribution of explanation contrastivity for a composer.

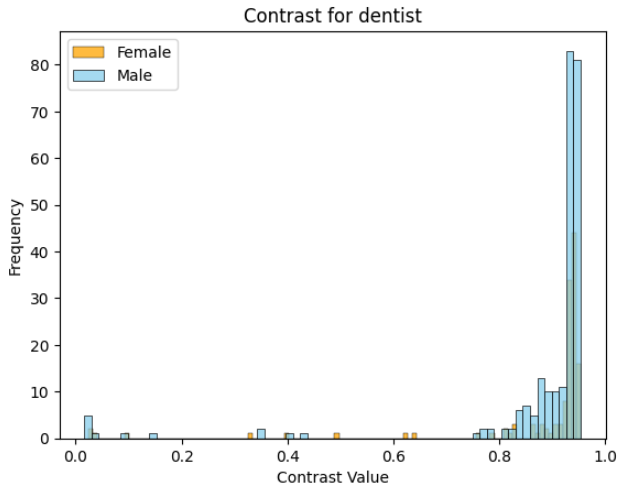

Figure B.6: The distribution of explanation contrastivity for a dentist.

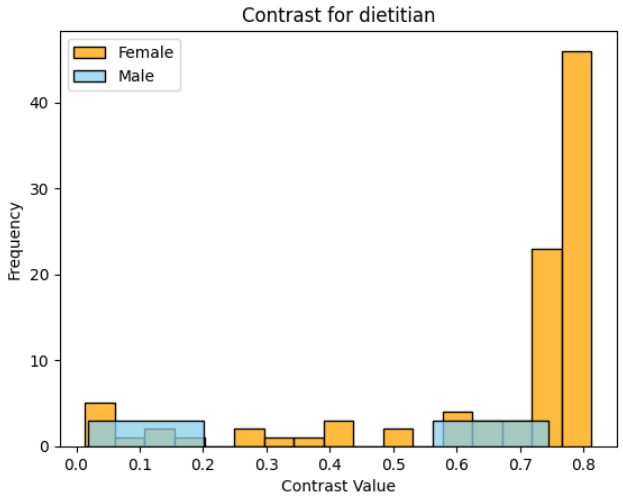

Figure B.7: The distribution of explanation contrastivity for a dietitian.

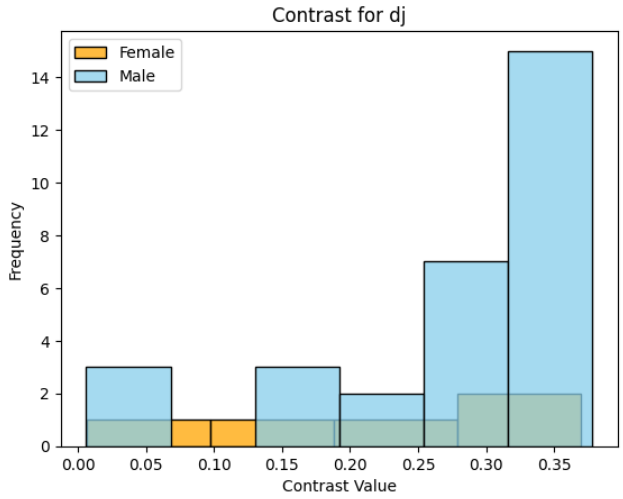

Figure B.8: The distribution of explanation contrastivity for a DJ.

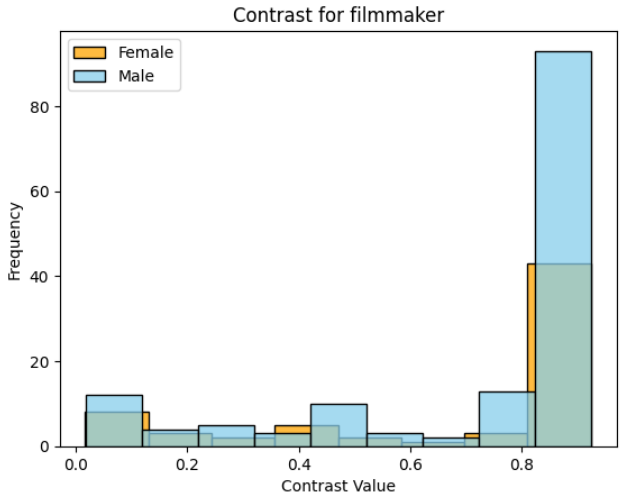

Figure B.9: The distribution of explanation contrastivity for a filmmaker.

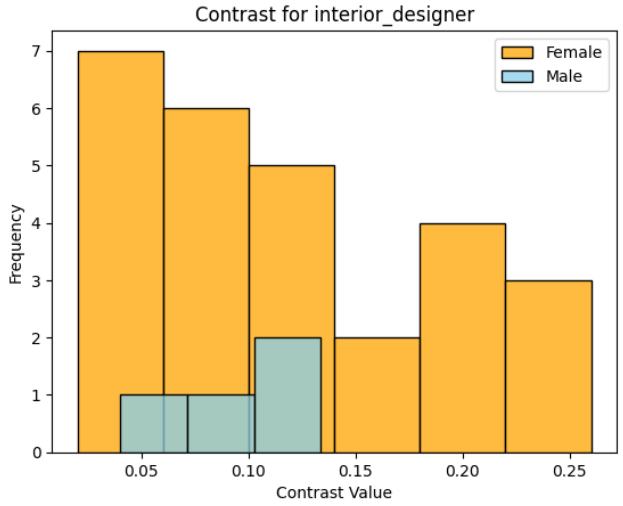

Figure B.10: The distribution of explanation contrastivity for an interior designer.

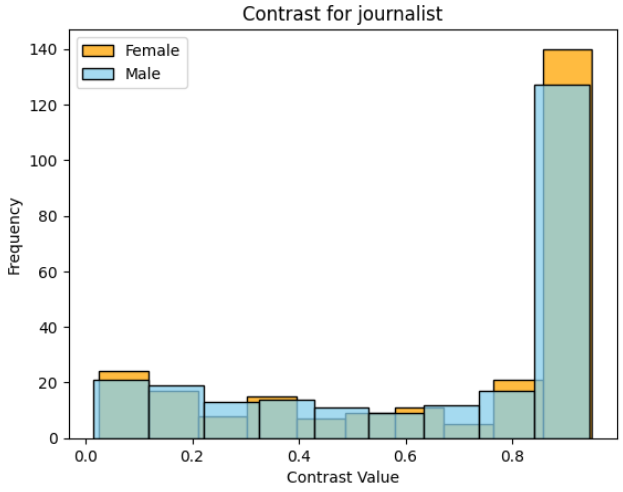

Figure B.11: The distribution of explanation contrastivity for a journalist.

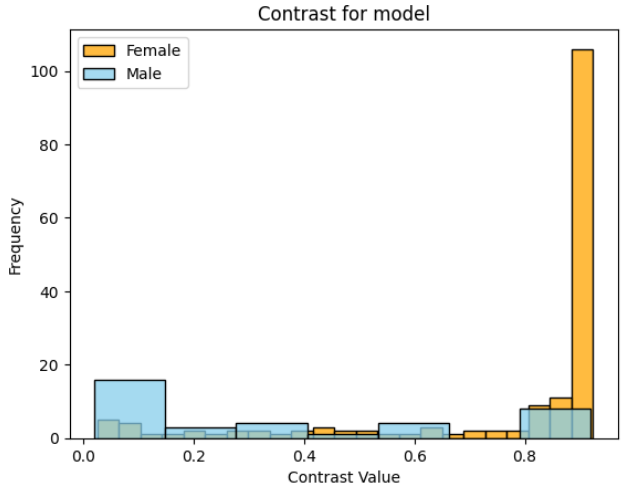

Figure B.12: The distribution of explanation contrastivity for a model.

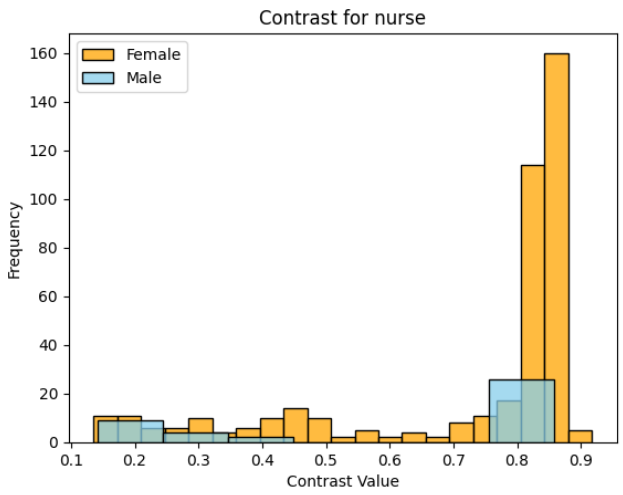

Figure B.13: The distribution of explanation contrastivity for a nurse.

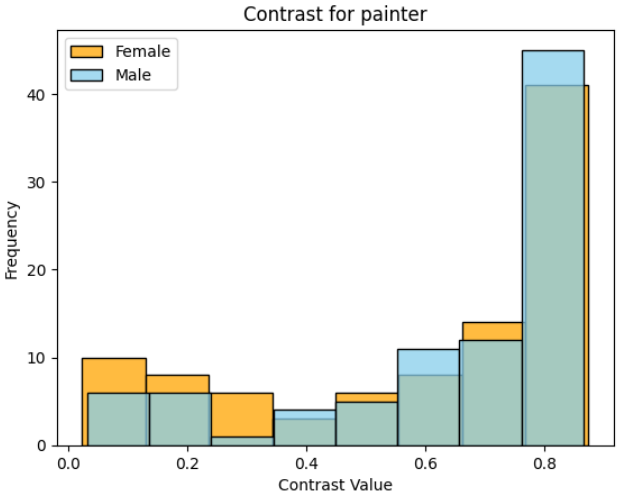

Figure B.14: The distribution of explanation contrastivity for a painter.

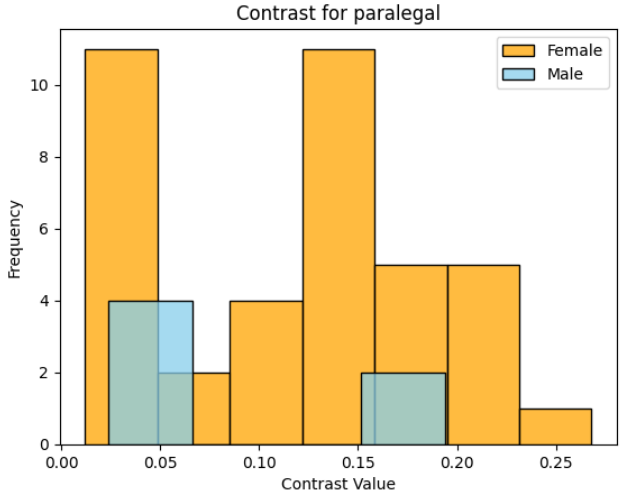

Figure B.15: The distribution of explanation contrastivity for a paralegal.

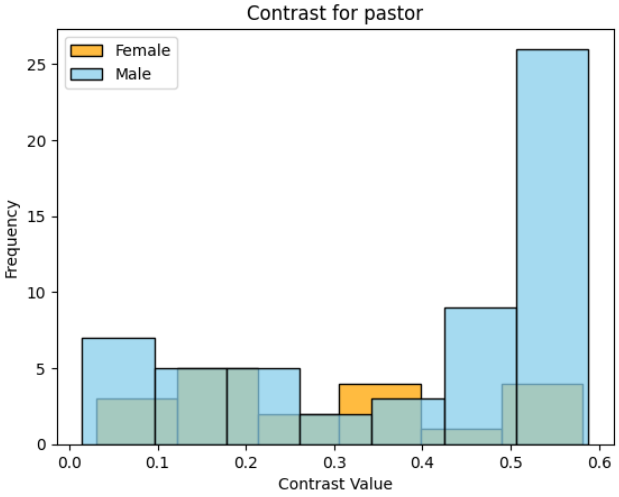

Figure B.16: The distribution of explanation contrastivity for a pastor.

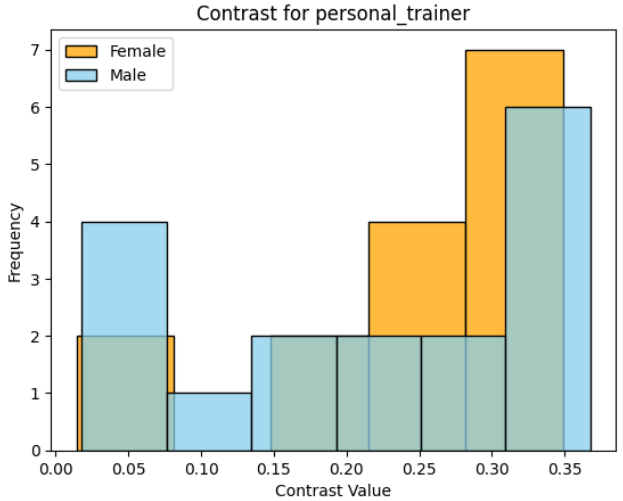

Figure B.17: The distribution of explanation contrastivity for a personal trainer.

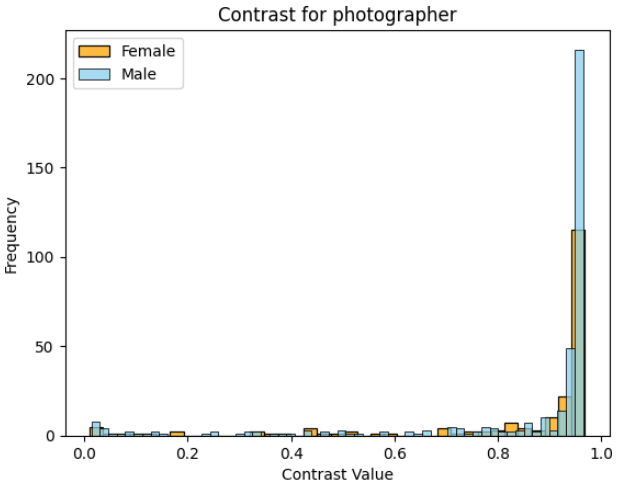

Figure B.18: The distribution of explanation contrastivity for a photographer.

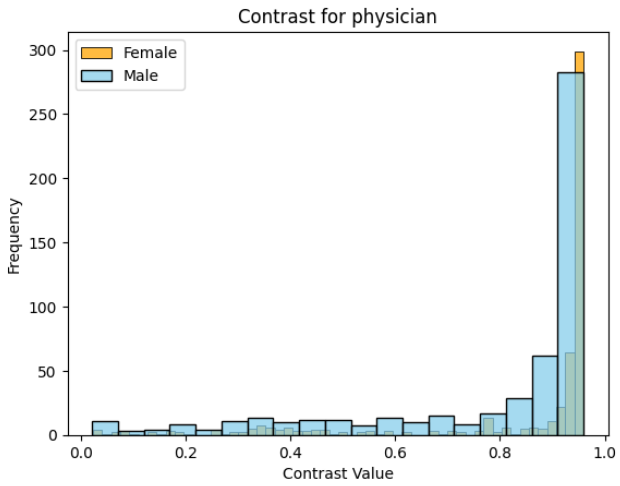

Figure B.19: The distribution of explanation contrastivity for a physician.

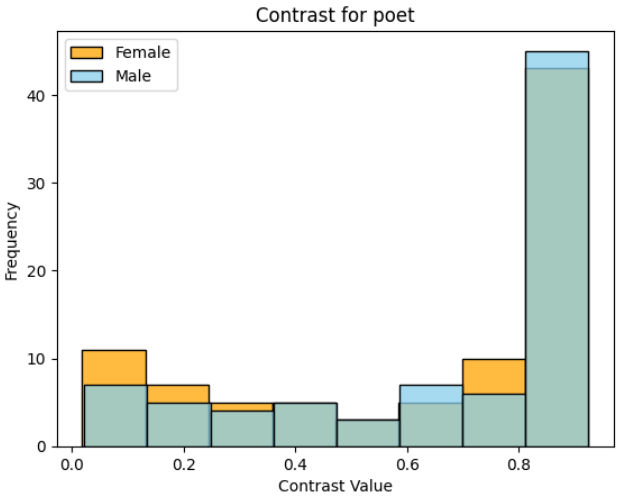

Figure B.20: The distribution of explanation contrastivity for a poet.

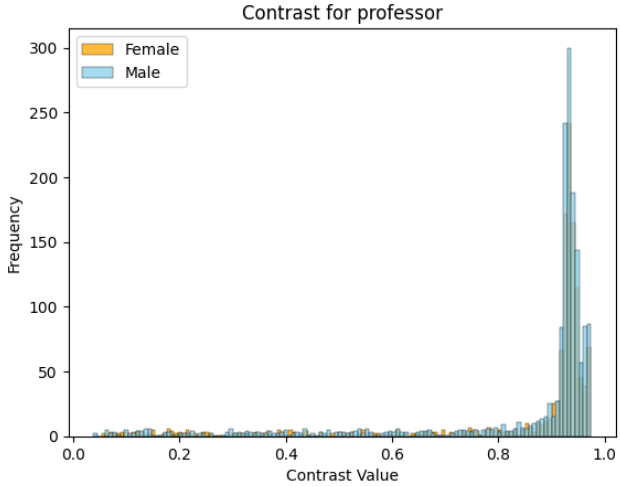

Figure B.21: The distribution of explanation contrastivity for a professor.

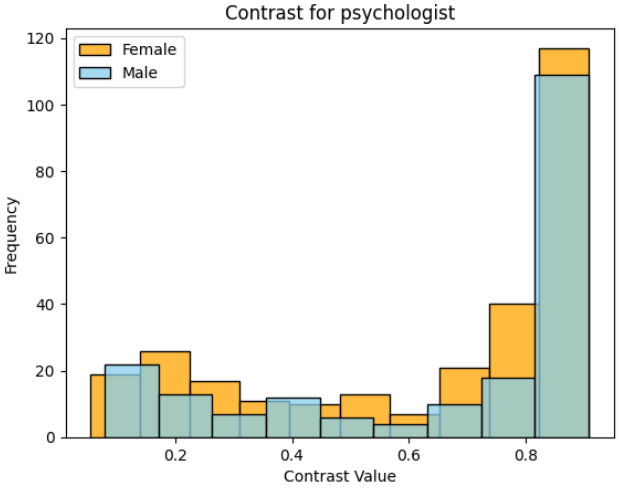

Figure B.22: The distribution of explanation contrastivity for a psychologist.

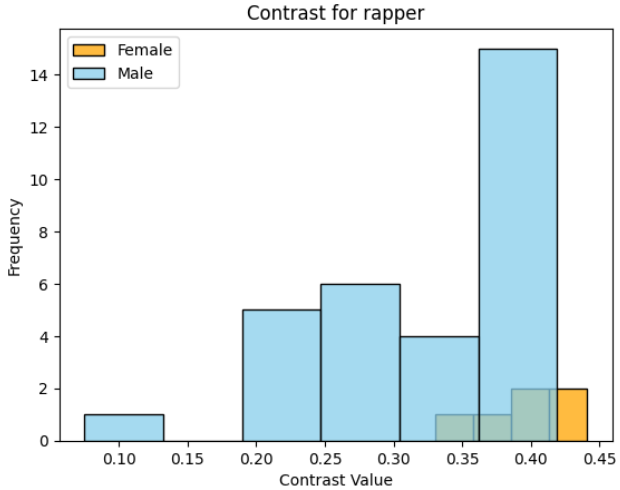

Figure B.23: The distribution of explanation contrastivity for a rapper.

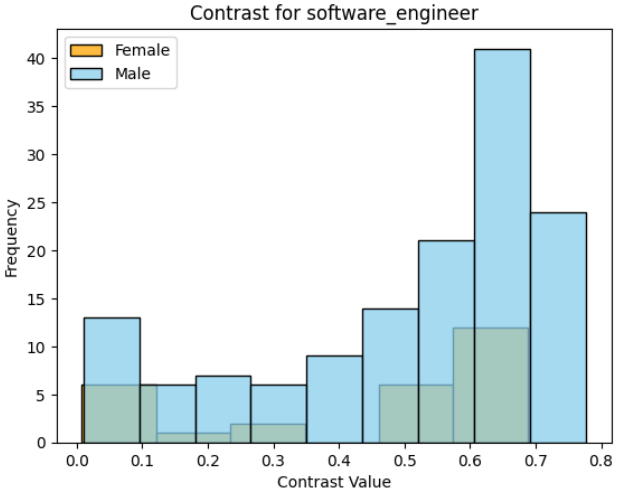

Figure B.24: The distribution of explanation contrastivity for a software engineer.

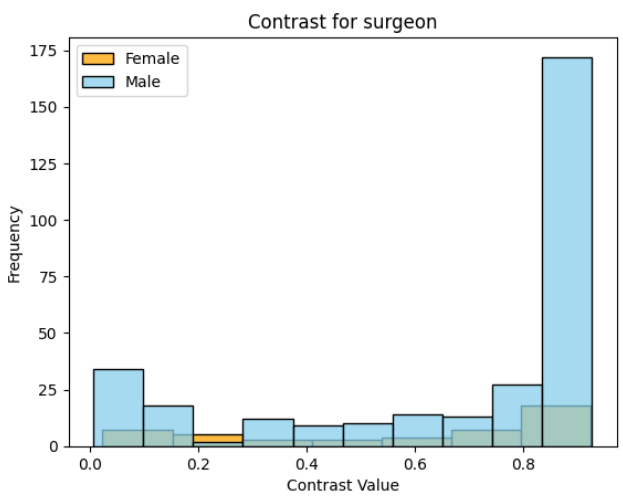

Figure B.25: The distribution of explanation contrastivity for a surgeon.

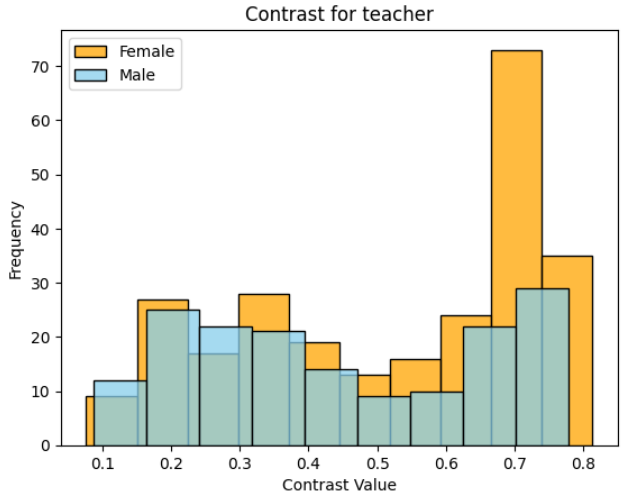

Figure B.26: The distribution of explanation contrastivity for a teacher.

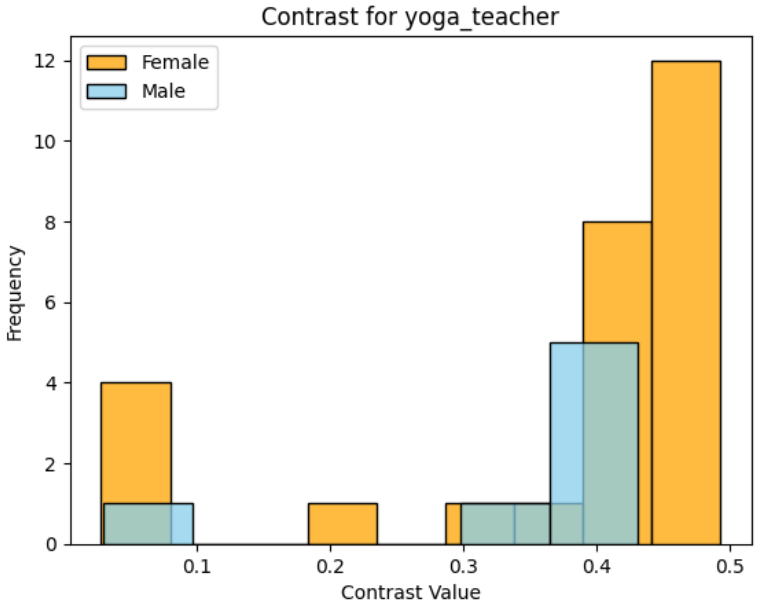

Figure B.27: The distribution of explanation contrastivity for a yoga teacher.

Appendix C

Code and Configuration

Listing C.1: Data loader used in experiments

```
def calc_contrast(sv, gender, pid):
    v = sv.values[:, pid]
    pos = np.abs(v[v > 0].sum())
    neg = np.abs(v[v < 0].sum())
    return max(pos, neg) / (1 + min(pos, neg))
```