DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
73

MARHARYTA DOMNICH

Advancing Human-Centric Counterfactual Explanations in Explainable AI

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
73

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
73

MARHARYTA DOMNICH

Advancing Human-Centric Counterfactual Explanations in Explainable AI

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on September 30, 2025 by the Council of the Institute of Computer Science, University of Tartu.

Supervisors

Prof. Dr. Raul Vicente Zafra
Institute of Computer Science
University of Tartu, Estonia

Dr. Eduard Barbu
Institute of Computer Science
University of Tartu, Estonia

Opponents

Prof. Barbara Hammer
Faculty of Technology
Bielefeld University, Germany

Assoc. Prof. Luca Longo
University College Cork, Ireland

The public defense will take place on January 15, 2026 at 12:15 in Narva Rd. 18-1018.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9908-57-081-5 (print) ISBN 978-9908-57-082-2 (pdf)

University of Tartu Press
http://www.tyk.ee/

To my family, friends, students, and colleagues
for making this journey possible and meaningful.

“І все на світі треба пережити,
і кожен фініш — це, по суті, старт.”

[“We have to live through everything,
and every finish is, in fact, a start.”]
— Lina Kostenko

ABSTRACT

Artificial Intelligence increasingly influences critical decisions across diverse domains like healthcare, education, and finance. The growing complexity and scale of these models often make their decision-making processes opaque, highlighting the importance of developing explanation methods that enhance transparency and accountability. The field of Explainable AI (XAI) aims to address this challenge by developing explanations that are meaningful to users. Human explanation processes are inherently complex and contrastive, often involving comparisons and hypothetical scenarios. This contrastive way of thinking is captured effectively by counterfactual explanations, which answer the question, "What minimal changes could alter a model's decision?". For counterfactual explanations to be effective, they must align closely with human preferences, ensuring they are meaningful, actionable, and trusted by users.

This thesis advances human-centric counterfactual explanations through four interconnected studies. By integrating insights from cognitive science, the research enhances both the generation and evaluation of counterfactual explanations across various domains.

The first study, inspired by human cognitive preferences, proposes the use of diffusion distance and directional coherence to enhance the search for counterfactual explanations. These innovations result in more feasible, human-centric explanations by emphasizing data connectivity and aligning changes in feature space with human reasoning patterns. Our approach, named Coherent Directional Counterfactual Explainer (CoDiCE), shows better performance in generating explanations that are both actionable and aligned with human explanatory virtues.

Addressing the critical issue of evaluating counterfactual explanations, the second study develops the CounterEval dataset, capturing detailed human judgments across multiple explanatory dimensions. Using data collected from over 200 participants, we introduce a unified evaluation framework that incorporates Large Language Models (LLMs) to predict averaged and individual human ratings, providing a scalable and consistent method to evaluate explanation quality. A subsequent analysis examines how perceived satisfaction with explanations can be modeled from other explanatory metrics (such as feasibility, trust, completeness, and complexity), providing deeper insights into the factors driving overall user satisfaction.

The practical impact of counterfactual explanations is further demonstrated in the context of medical imaging by introducing a COunterfactual INpainting approach (COIN) for weakly supervised semantic segmentation in medical imaging. COIN generates explanations by flipping classification outcomes from abnormal to normal, using the differences between the original and altered images as weak segmentation labels. Applied to kidney tumor segmentation, this methodology significantly reduces the manual labeling workload for radiologists and enables pathology segmentation in scenarios lacking extensively annotated datasets. Counterfac-

tual inpainting significantly outperforms attribution-based methods, showcasing the real-world potential of counterfactual explanations in healthcare.

Together, these studies contribute to the field of XAI by developing and validating counterfactual explanation methods that enhance the transparency and usability of AI systems and also closely align with human cognitive processes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence 17, 21–24, 39, 66, 76
ALTAI	The Assessment List for Trustworthy Artificial Intelligence 23
CAM	Class Activation Map 11, 65, 68
CAV	Concept Activation Vector 35
CCE	Conceptual Counterfactual Explanations 35
cGAN	conditional Generative Adversarial Network 35
CoDiCE	Coherent Directional Counterfactual Explainer 6, 45
COIN	COnterfactual INpainting 6, 10, 20, 65–70
CT	Computed Tomography 20, 65, 66
CV	Counterfactual Validity 11, 68, 69
DiCE	Diverse Counterfactual Explanations 30, 45
FACE	Feasible and Actionable Counterfactual Explanations 40, 45
FID	Fréchet Inception Distance 11, 68, 69
GAN	Generative Adversarial Network 10, 35, 66, 67, 70
Grad-CAM	Gradient-weighted Class Activation Mapping 22, 27, 28
IoU	Intersection over Union 69
LIME	Local Interpretable Model-Agnostic Explanations 22, 25–28
LLM	Large Language Model 6, 19, 51, 60, 64, 75, 76
LRP	Layer-wise Relevance Propagation 22
MAD	Median Absolute Deviation 37
MIP	Mixed-Integer Programming 33
ML	Machine Learning 25
MSE	Median Squared Error 31

RQ Research Question 18, 40

SAE Sparse Autoencoders 28
SHAP SHapley Additive exPlanations 22, 26–28

TRUST-AI Transparent Reliable and Unbiased Smart Tool 50
TUH Tartu University Hospital 10, 11, 65, 68, 69
TV Total-Variation 67

VAE Variational Autoencoders 35

WSSS Weakly Supervised Semantic Segmentation 65, 66

XAI Explainable AI 6, 17, 18, 21–23, 30, 40, 63

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I. **Domnich**, **Marharyta** and Vicente, Raul (2024). "Enhancing counterfactual explanation search with diffusion distance and directional coherence". In: *Explainable Artificial Intelligence*. Vol. 2155. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, pp. 60–84. DOI: 10.1007/978-3-031-63800-8_4.

II. **Domnich**, **Marharyta**, Välja, Julius, Veski, Rasmus Moorits, Magnifico, Giacomo, Tulver, Kadi, Barbu, Eduard, and Vicente, Raul (2025a). "Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 15, pp. 16308–16316. DOI: 10.1609/aaai.v39i15.33791.

III. **Domnich**, **Marharyta**, Veski, Rasmus Moorits, Välja, Julius, Tulver, Kadi, and Vicente, Raul (2025b). "Predicting Satisfaction of Counterfactual Explanations from Human Ratings of Explanatory Qualities". In: *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland, pp. 210–229. DOI: 10.1007/978-3-032-08317-3_10.

IV. Shvetsov, Dmytro, Ariva, Joonas, **Domnich**, **Marharyta**, Vicente, Raul, and Fishman, Dmytro (2024). "COIN: Counterfactual inpainting for weakly supervised semantic segmentation for medical images". In: *Explainable Artificial Intelligence*. Vol. 2155. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, pp. 39–59. DOI: 10.1007/978-3-031-63800-8_3.

Publications not included in the thesis

I. Sakkas, Nikos, Yfanti, Sofia, Daskalakis, Costas, Barbu, Eduard, and **Domnich**, **Marharyta** (2021). "Interpretable forecasting of energy demand in the residential sector". In: *Energies* 14.20. ISSN: 1996-1073. DOI: 10.3390/en14206568.

II. Sakkas, Nikos, Yfanti, Sofia, Shah, Pooja, Sakkas, Nikitas, Chaniotakis, Christina, Daskalakis, Costas, Barbu, Eduard, and **Domnich**, **Marharyta** (2023). "Explainable approaches for forecasting building electricity consumption". In: *Energies* 16.20. ISSN: 1996-1073. DOI: 10.3390/en16207210.

III. Majoral, Daniel and **Domnich**, **Marharyta** (May 2025). "Kaizen: Decomposing cellular images with VQ-VAE". In: *PLOS ONE* 20.5, pp. 1–11. DOI: 10.1371/journal.pone.0313549.

IV. Khajuria, Tarun, Dias, Braian Olmiro, **Domnich**, **Marharyta**, and Aru, Jaan (2025). "Interpreting the structure of multi-object representations in vision encoders". In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 359–382. DOI: 10.1007/978-3-032-08324-1_16.

V. Barbu, Eduard, **Domnich**, **Marharyta**, Vicente, Raul, Sakkas, Nikos, and Morim, André (2025). "Exploring Commonalities in Explanation Frameworks: A Multi-Domain Survey Analysis". In: arXiv: 2405.11958 [cs.LG].

VI. **Domnich**, **Marharyta**, Sünter, Indrek, Trofimov, Heido, Wold, Olga, Harun, Fariha, Kostiukhin, Anton, Järveoja, Mihkel, Veske, Mihkel, Tamm, Tanel, Voormansik, Kaupo, et al. (2021). "KappaMask: AI-based cloudmask processor for Sentinel-2". In: *Remote Sensing* 13.20. ISSN: 2072-4292. DOI: 10.3390/rs13204100.

Author's contribution to the publications

In Publication I, the author was responsible for all stages of the research, including formulating the idea and hypotheses, developing the methodology, conducting experiments, and writing the publication.

In Publication II, III, the author developed research ideas and hypotheses, collaborated on designing the methodology, contributed to the experiments, and took the leading role in writing both publications.

In Publication IV, the author developed the research idea, formulated hypotheses, contributed to improving the algorithm and methodology from the Explainable AI perspective, co-supervised a student conducting experiments, and took a leading role in writing the manuscript.

PREFACE

PhD research is a journey of personal transformation. Every transformation is unique, starting from different starting points and following different trajectories. I would like to share some observations with my fellow PhD colleagues reading this, perhaps you will find that you are not alone in your struggle.

Researching in the field often feels like *climbing a ladder* of several stages.

Stage I - Curious Beginnings. The first step of this ladder is filled with motivation and hope. You are driven by curiosity, eager to explore, and convinced that answers are within reach. Every paper seems fascinating, opening a new door.

Stage II - Critical Awakening. As you delve deeper into your field, the cracks begin to appear: gaps, contradictions, problems, and assumptions built upon other assumptions. Suddenly, nothing seems to make complete sense. Curiosity gives way to skepticism. It is an exhausting stage, and probably the one where you spend the most time during your PhD.

Stage III - Informed Curiosity. Then comes the realization that everyone is, in some way, as lost as you are. Once you have read enough and accepted the reality of your field, you begin to see that imperfection is not a flaw but a possibility. Curiosity returns, this time grounded in realism. Despite the gaps, you recognize the standards of your field and decide to push one boundary at a time. And when you do, it feels deeply rewarding, reminding you why you started this journey in the first place. With that, new hope returns.

I will live to learn if it is a ladder or a circle ...

1. INTRODUCTION

In the last decade, Artificial Intelligence (AI) has been reshaping every domain, including those that influence critical decisions for society, such as healthcare, finance, criminal justice, and education. From automated loan approvals to tumor segmentation in medical imaging, predictive models are increasingly entrusted with decisions that carry significant personal and societal consequences (Obermeyer et al. 2019; Rudin 2019). However, the performance of such models that have remarkable predictive capabilities is attributable primarily to their complexity and the vast amount of data they consume, factors that inherently make such models opaque (Gunning et al. 2019). This opacity poses a significant challenge for the field of Explainable AI (XAI), creating a pressing need for methods that clarify not only what decision was made but also why it was made and how an unfavorable outcome might be altered. Although simpler, intrinsically interpretable models exist, they often lag behind more complex architectures in predictive accuracy. This reality underscores the necessity of post-hoc explanation methods that preserve high performance while offering transparent reasoning. Since compromising the complexity of the model would result in its insufficient performance (Lipton 2018), post-hoc explanation methods emerged as a practical compromise. Such methods include local surrogate models (Ribeiro et al. 2016), feature-importance approaches (Lundberg et al. 2017), counterfactual explanations (Wachter et al. 2017), and others with each addressing the transparency gap in different ways.

One particularly compelling post-hoc explanation method is ***counterfactual explanations***. Rather than presenting feature importances or visual attributions, counterfactual explanations answer the specific question: "What minimal changes to the input would have flipped the model's decision?" (Wachter et al. 2017). This approach resonates strongly with human cognition, as people naturally engage in "what if" counterfactual thinking to reflect on how different actions might yield different outcomes (Miller 2019). For example, consider a loan application scenario in which a loan application is denied. A counterfactual explanation might indicate that if the applicant's annual income were increased by 5%, or if the level of education had the value of "PhD" instead of "Master's", the decision could have been reversed to approval. Such explanations not only highlight the factors that influenced the decision but also provide actionable guidance, helping individuals to understand which specific changes might yield a more favorable outcome (Poyiadzi et al. 2020). In a loan application context, for instance, a user may initially suspect discriminatory factors such as nationality or gender when a loan is denied. Humans often attribute unfavorable outcomes to potentially discriminatory factors, influenced by negativity bias (Kuhl et al. 2023). However, a counterfactual explanation indicating that a modest increase in salary or a slightly larger savings balance could have shifted the decision from "deny" to "approve" effectively draws attention to feasible, actionable changes. By directly linking a decision to minimal changes in input features, counterfactual explanations foster a sense of control and

understanding among end users, serving as a bridge between the precision of algorithmic outputs and the qualitative judgments that humans naturally employ when assessing causality and responsibility (Pearl 2009).

The rationale behind counterfactual explanations is deeply rooted in theories of causality. It is important to note that counterfactual explanations in Explainable AI are not equivalent to counterfactuals in causal inference. In the realm of causal inference, counterfactuals are used to answer questions about how different actions or conditions might change outcomes under a given set of assumptions, representing the highest level of Pearl's "ladder of causality" (Pearl 2019). In contrast, counterfactual explanations in XAI typically focus on exploring local decision boundaries of a trained model in a post-hoc way without requiring full causal models. This practical, model-centered focus does not necessarily account for genuine causal relationships in the real world. Nevertheless, a nuanced understanding of causality can enrich these explanations by ensuring that the hypothetical changes suggested are both plausible and informative. Indeed, the term "counterfactual" itself originates in causal inference, which sometimes causes confusion. However, within XAI, 'counterfactual' refers primarily to minimal local changes that alter a model's prediction, rather than a rigorous causal intervention.

In addition to causal reasoning, cognitive science provides insights into the biases and heuristics that influence human judgment. It is important to recognize that for any given point, multiple counterfactual explanations can be generated by sampling around the decision boundary. Consider the famous anecdote involving bank robber Willie Sutton. When asked by the journalist, "Why did you rob the bank?", he allegedly answered, "Because that's where the money is". This answer highlights the assumption about different contrasts that the journalist had in mind: "Why did you rob the bank rather than not robbing at all?" versus "Why did you rob the bank rather than another place, i.e., university?". The answer would not be considered a good explanation for multiple reasons. First, the expected contrast is not fulfilled by the answer. Second, the answer restates what is already widely understood and does not provide new knowledge. While mathematically valid counterfactual explanations may exist for a given decision, human evaluators tend to prefer those that are contextually meaningful. Research indicates that individuals are more likely to value explanations that suggest small, incremental changes than those that propose unrealistic or radical adjustments (Kahneman 2011). On the other side, Zemla et al. (2017) suggests that rather than adding small perturbations to multiple features, humans tend to prefer larger changes in a few key features, creating a trade-off between sparsity and feasibility. Therefore, accounting for human cognitive preferences is an important consideration that informs the development of explanation methods throughout this thesis.

This thesis aims to enhance the effectiveness, cognitive alignment, evaluation, and applicability of counterfactual explanations guided by the following research questions (RQ):

- **RQ1**: Can we extract more feasible and coherent explanations compared to

existing methods?

- **RQ2a**: Can overall satisfaction with counterfactual explanations be predicted using human ratings of explanatory metrics? Specifically, what is the influence of demographic data on ranking patterns, and which explanatory qualities contribute most to overall satisfaction?
- **RQ2b**: Can large language models (LLMs) reliably estimate the quality of counterfactual explanations by mimicking human judgment?
- **RQ3**: Can we apply counterfactual explanations to produce Weakly Supervised Semantic Segmentation labels effectively?

The thesis is structured around four primary investigations that collectively aim to advance the state of the art in counterfactual explanations. First, in Chapter II, we review the literature on explainable AI, counterfactual explanations, causal inference, and cognitive science, setting the theoretical foundations for the work presented in subsequent chapters. Next, in Chapter III, we address RQ1 by translating key human preferences, in particular, feasibility and coherence, into mathematical formulations using diffusion distance and directional coherence. These measures are integrated into the objective function for the counterfactual search process, aiming to produce explanations that are both algorithmically faithful and consistent with human intuition. [Publication 1 ***Enhancing Counterfactual Explanation Search with Diffusion Distance and Directional Coherence*** (**Domnich** et al. 2024)].

Expanding on the understanding of human preferences in counterfactual explanations, Chapter IV addresses RQ2a and RQ2b by tackling the critical challenge of evaluation. Recognizing that quantitative metrics of counterfactual explanation evaluation often fail to capture human preferences fully and that traditional user studies face scalability issues, we compile a list of important evaluation metrics, construct a diverse dataset of counterfactual scenarios covering ranges of these metrics, and collect human judgments across those evaluation metrics. In parallel, we investigate which measured biases contribute to forming overall satisfaction, where we highlight the importance of feasibility and trust of counterfactual explanations, but also completeness, consistency, fairness, and complexity when dealing with specialized user groups and applications. [See Publication 2 ***Predicting Satisfaction of Counterfactual Explanations from Human Ratings of Explanatory Qualities*** (**Domnich** et al. 2025b)]. Additionally, given emerging capabilities of Large Language Models (LLMs), we assess their potential to become scalable proxies for human evaluative judgments. [See Publication 3 ***Towards Unifying Evaluation of Counterfactual Explanations: Leveraging Large Language Models for Human-Centric Assessments*** (**Domnich** et al. 2025a)].

Finally, Chapter V addresses RQ3 by demonstrating the applicability of our methods to complex, high-stakes domains such as medical imaging, extending beyond tabular or structured data. We investigate the potential to apply counterfactual explanations to more complex domains such as medical imaging, where limited

labeled data and high stakes further amplify the need for intuitive, actionable explanations. Medical imaging is particularly critical because it involves diagnosing or identifying potential health risks from high-dimensional data, where each pixel may carry clinically meaningful information. Additionally, the limited availability of labeled data, together with the severe consequences of misdiagnosis, brings the need for interpretable explanations that help to gain the trust of clinicians in automation. Tackling such complex, real-world applications, we demonstrate that counterfactual explanations can help beyond tabular scenarios and classification problems. Therefore, guided by the intuition that a good explanation of a classifier should also serve as a segmentation mask, we developed the COnterfactual INpainting (COIN) algorithm. When a classifier labels a Computed Tomography (CT) scan as "abnormal", this counterfactual approach trains a generator to produce a corresponding "normal" image that flips the classifier's prediction. The difference between the original and counterfactual images not only effectively explains the classifier's decision but can also serve as a weak segmentation mask, thereby saving radiologists' resources. [See Publication 4 ***Counterfactual inpainting for weakly supervised semantic segmentation for medical images*** (Shvetsov et al. 2024)]

The final chapter summarizes key findings, discusses limitations, and outlines potential directions for future research. Collectively, the four interconnected studies presented in three chapters of this thesis contribute to improving algorithms for generating counterfactual explanations by incorporating human cognitive preferences, standardizing and scaling their evaluation from a human-centered perspective, and promoting their adoption beyond the field of Explainable AI.

2. BACKGROUND

2.1. Introduction to Explainable AI

The quest for Explainable AI can be traced back to the era of expert systems in the 1970s and 1980s (Clinciu et al. 2019). For example, systems like MYCIN could justify their diagnoses via human-readable rules created by domain experts, outlining their reasoning process. These symbolic AI systems were *inherently interpretable*, as their decision logic (if-then rules) was transparent by design. At the time, the importance of explaining the chain of reasoning behind AI decisions was already recognized. Classical AI models, including decision trees and linear regression models, provided built-in transparency, either by giving the decision pathways or coefficients. Such early interpretability methods illustrated that transparency and functionality could coexist within AI systems, laying crucial foundations for modern XAI.

Before the formal rise of the XAI field, researchers developed early methods aimed at interpreting complex models. Techniques such as rule extraction in the 1990s attempted to distill neural network decisions into comprehensible if-then rules. At the same time, sensitivity analysis provided insights into the influence of small input perturbations on model outputs. Despite being innovative, these methods struggled to scale effectively to the increased complexity of deep neural architectures. They were predominantly model-specific, limiting their practical adoption in different AI models and contexts (Adadi et al. 2018). Consequently, these early approaches often failed to faithfully represent the underlying complex decision-making processes, motivating the need for more universally applicable and effective interpretability tools.

However, the early 21st century saw a shift towards more complex *black-box* models, such as deep neural networks and ensemble methods, prioritizing accuracy at the cost of interpretability (Vilone et al. 2021). These advanced methods excelled in predictive tasks across various domains, but failed to inherently provide explanations for their decisions, creating what is now known as the "black-box" problem. Consequently, the trade-off between accuracy and explainability became central to XAI research (Adadi et al. 2018; Arrieta et al. 2020). The opacity of the models hindered trust and accountability, especially in critical fields such as healthcare, finance, and autonomous systems, where the rationale behind decisions significantly impacts human life and creates interest in *post-hoc* explanation techniques.

A significant moment came with the launch of DARPA's XAI program in 2016, which invested in new techniques that produce explainable yet high-performance models, called "glass box", and explanation interfaces (Gunning et al. 2019). The DARPA program (2017–2021) formally put Explainable AI on the map, funding numerous projects to develop AI systems that are both high-performing and understandable by end-users. Around the same time, influential algorithms were in-

troduced, such as the introduction of LIME (Local Interpretable Model-Agnostic Explanations) in 2016, a technique to explain any classifier's prediction by approximating it locally with an interpretable model (Ribeiro et al. 2016). The following year, SHAP (SHapley Additive exPlanations) was proposed to attribute feature importance for any model grounded in game theory (Lundberg et al. 2017). Together, LIME and SHAP became widely used tools, marking the transition of XAI from concept to practical toolkit. In parallel, techniques like DeepLIFT (Shrikumar et al. 2017), Integrated Gradients (Sundararajan et al. 2017), Grad-CAM (Selvaraju et al. 2017), RISE (Petsiuk et al. 2018), and Layer-wise Relevance Propagation (LRP) (Bach et al. 2015) emerged for explaining deep neural networks' predictions, especially in computer vision.

With increased interest in XAI, dedicated workshops and conferences, including ACM FAccT (Fairness, Accountability, and Transparency), ICML workshops on interpretability began highlighting explainability alongside ethics, reflecting how XAI evolved from scattered interpretability efforts into a cohesive field motivated by the dual need to open the black box of AI and meet societal demands for transparency. Latest conferences dedicated to XAI, such as the World Conference on eXplainable Artificial Intelligence, highlight the multidisciplinary nature of the field and the need to draw insights from the literature of philosophy, psychology, and cognitive science (Thommes et al. 2024; Vilone et al. 2023).

Multiple factors contribute to the trend of post-hoc explanation techniques: the need for transparency in high-stakes decisions, trust and user acceptance of AI, and emerging legal regulations. For instance, the EU GDPR stated a "right to explanation," pressuring organizations to make automated decisions more transparent (European Parliament and Council of the European Union 2016). In response, scientists and policymakers began calling on the community to frame explainability in scientific terms. For instance, Doshi-Velez et al. 2017 emphasized the lack of formal definitions and evaluation for interpretability. Similarly, Lipton 2018 critiqued the loose use of terms in the "Mythos of model interpretability", drawing attention to the many meanings of interpretability and the need to distinguish them clearly.

The upcoming EU AI Act proposals have pressured companies to provide meaningful information about automated decisions. The EU AI Act introduces a risk-based framework with strict requirements for "high-risk" AI systems, many of which directly address explainability and transparency (European Commission 2023). High-risk AI (e.g., algorithms for credit scoring, hiring, medical devices, etc.) must be sufficiently transparent so that their outputs can be understood and appropriately used. For example, Article 13 of the Act requires clear user instructions that "explain how to interpret the system's output" and describe the system's capabilities, limitations, and potential risks. These instructions should include details like the model's intended purpose, its accuracy metrics, and "information that is relevant to explain its output" in understandable terms (European Commission 2023, Article 13). Beyond technical documentation, the Act mandates trans-

parency disclosures in user interaction, meaning the system must notify people that they are dealing with a machine, and generative AI content must be clearly labeled as AI-generated (European Commission 2023, Article 52). The United States has likewise highlighted explainability, for example, through the 2023 Executive Order on trustworthy AI, which emphasizes transparency in AI use (The White House 2023). In specific sectors like finance (credit scoring) and healthcare, regulations now mandate that decisions are explainable to end-users or auditors (Wachter et al. 2017). Ethically, XAI is tied to the principles of fairness and non-discrimination. Opaque models can mask biased decision criteria, so explainability can be seen as a tool to detect and mitigate bias (Mehrabi et al. 2021). For instance, counterfactual explanations have been proposed as a way to explain decisions in understandable, human terms (e.g., explaining loan rejections by stating what changes would result in approval) and to test models for unfair bias (Karimi et al. 2022). In summary, upcoming laws and ethical guidelines are forcing AI systems to be more transparent, making XAI research not only academically relevant but essential for compliance and responsible AI deployment (Arrieta et al. 2020).

Nowadays, XAI has evolved into a mainstream concern in AI research and deployment. Companies incorporate explainability dashboards into AI services, and regulators draft guidelines emphasizing transparency and trustworthiness in AI deployment (Artificial Intelligence (AI HLEG) 2019, The Assessment List for Trustworthy Artificial Intelligence (ALTAI)). Research has expanded beyond initial feature-attribution methods to broader explanation modalities (natural language explanations, visualizations, interactive explanations, etc.). Companies such as Google, IBM, and Microsoft have incorporated explainability toolkits (e.g., Google's Explainable AI service (Google 2023), IBM's AI Explainability 360 (Arya et al. 2019), and Microsoft's InterpretML (Nori et al. 2019)) to help users interpret model outputs. Importantly, the community began exploring evaluation metrics for explanations and the human side of interpretability. The field has grown so much that Schwalbe et al. (2024) conducted a "survey of surveys" to unify the myriad XAI taxonomies.

Therefore, the broad adoption of XAI signifies a cultural shift in AI development: Success is no longer measured only by accuracy, but also by how well humans can interpret and trust the reasoning of the model.

2.2. Taxonomy of Explainable AI

In this subsection, we provide definitions of the key terminology used throughout this thesis. Although we acknowledge that this overview is not exhaustive, we aim to clarify critical terminology to aid in understanding our research.

2.2.1. Explainability and Interpretability definitions

The definitions of explainability and interpretability have evolved over time, remaining still conflicting. Researchers initially used terms like transparency, in-

telligibility, interpretability, and explainability, loosely and often interchangeably (Clinciu et al. 2019). Lipton described interpretability as an "important and slippery" concept, noting that many papers proclaimed models "interpretable" without defining the term (Lipton 2018). Similarly, (Doshi-Velez et al. 2017) noted there is minimal consensus on what interpretable machine learning is and how it should be measured, urging for a more rigorous approach. They defined interpretability as "the ability to explain or to provide the meaning in understandable terms to a human", and explainability as "an interface between humans and the decision-maker (AI) that produces details or reasons for its functioning". Nonetheless, the historical trend shows increasing awareness that precise terminology matters (Arrieta et al. 2020). While a decade ago, authors might use interpretability/explainability as self-evident words, today's researchers are more careful to define these terms or at least acknowledge the lack of a single consensus definition.

The academic literature reveals two broad camps on this issue (Graziani et al. 2023). In one camp, researchers use interpretability and explainability almost synonymously without sharply differentiating the two. They treat both as umbrella terms for making models understandable, either intentionally or unintentionally, primarily when referring to human insight into the model without digging into technical nuance (Adadi et al. 2018; Miller 2019). Miller uses a definition from Biran and Cotton that interpretability is "the degree to which an observer can understand the cause of a decision". In this view, an explanation is a mechanism to achieve interpretability, acting as a means of communicating the cause of a decision in an understandable way.

In the other camp, scientists draw a clear distinction between interpretability and explainability (Arrieta et al. 2020; Lipton 2018; Longo et al. 2024; Rudin 2019). A common differentiation is that interpretability describes an intrinsic property of the model (often implying a transparent or simple model). In contrast, explainability refers to a post-hoc capability to generate an explanation for any model, including black-box models. In this view, an interpretable model might be something inherently transparent like a small decision tree (no external explanation needed). At the same time, an explainable approach might take a complex neural network and produce an explanation (e.g., a feature importance plot or a textual justification) to help a user understand its output. Rudin (2019) advocates strongly for this distinction, arguing that we should design interpretable models directly, rather than relying on Explainable AI techniques to explain opaque models. Arrieta et al. (2020) differentiate the terms as noted above: interpretability is the ability of the model to be understood. At the same time, explainability is a property of the tool or interface that communicates the model's reasoning. These nuances have practical implications. If one treats explainability and interpretability as the same, one might focus on any method that improves human understanding. If one sees them as distinct, one might prioritize intrinsic interpretability (through model choice) vs. post-hoc explainability (through explanatory techniques) differently. The debate is ongoing, but the differentiated view is increasingly popular

among researchers, explicitly stating their stance. The lack of consensus on terminology has led to efforts, such as Sokol et al. (2020)'s work to catalog the properties of explanations systematically. Sokol et al. (2020) introduced Explainability Fact Sheets that enumerate dimensions along which an XAI method can be characterized. However, they acknowledge that explanations in ML "come in many forms, but a consensus regarding their desired properties is yet to emerge". Some explicitly define interpretability as model transparency and explainability as the ability to provide explanations. In the recent manifesto (Longo et al. 2024), the researchers urge against inventing new definitions, but to formalize existing definitions and standardize terminology.

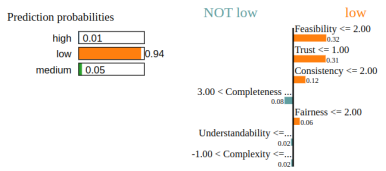
This thesis follows Rudin and Lipton's definitions that **interpretability** relates to a model's inherent understandability. In contrast, **explainability** refers to the ability to generate explanations for a model's outputs, stating that any interchangeable term used is accidental. Furthermore, **transparency** refers to the intrinsic interpretability of the model itself, whose structure and parameters are understandable, often described as *white-box*. Transparency can be further decomposed into aspects like simulatability (a person could reason through the model step by step), decomposability (each part of the model has an intuitive explanation), and algorithmic transparency (understanding the training process) (Lipton 2018). Interpretability in this sense is an inherent property of the model. We note the existence of partial interpretability, when the complex model architecture is disentangled or has a structured representation (Khajuria et al. 2025). In this case, the model is often called *grey-box* when the entire model is not interpretable, but parts of it have interpretable qualities.

2.2.2. Post-hoc and Ante-hoc Methods

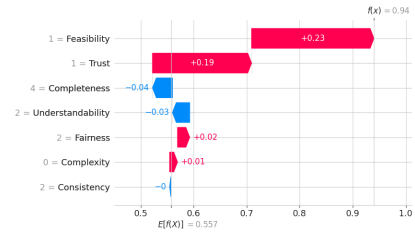
This distinction concerns when the explanation is generated relative to model training. **Ante-hoc** (intrinsic) methods are "interpretable by design" when the model itself is constrained to be transparent or self-explanatory (Molnar 2020). Examples include decision trees, linear models, or models that output human-readable logic along with predictions. In contrast, **post-hoc** methods generate explanations after a complex model is trained, without altering that model. Post-hoc explainers function as external analysis tools applied to pre-existing "black-box" models, interpreting their behavior post-training. For instance, LIME is a post-hoc method that creates explanations for individual predictions of any trained model. In contrast, a decision tree is an ante-hoc model whose interpretable structure is the explanation itself. Ante-hoc approaches often sacrifice some accuracy for interpretability, whereas post-hoc approaches approximate explanations for highly accurate black-box models.

2.2.3. Local and Global Explanations

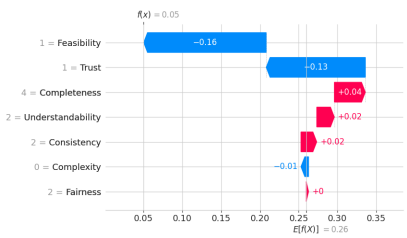
Explainability methods also differ in their scope of application. **Local explanations** provide insights about specific predictions or particular instances, addressing questions like "Why did the model make this decision for this particular case?". For instance, a local explainer might highlight the features of a specific loan application that led an AI model to reject it. LIME and SHAP are prototypical local explainers that can provide detailed, instance-level insights. For instance, Figure 1 illustrates explanations for the prediction of satisfaction with counterfactual explanations classified into three categories: high, medium, and low, based on their explanatory qualities. In contrast, **global explanations** aim to clarify the model's



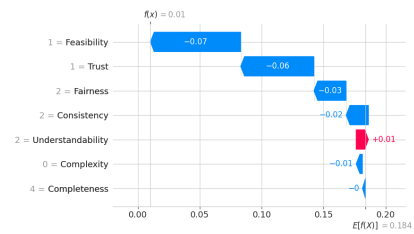
(a) LIME explanation illustrating the predicted probability of "low satisfaction".



(b) SHAP waterfall plot for predicting "high satisfaction".



(c) SHAP waterfall plot for predicting "medium satisfaction".



(d) SHAP waterfall plot for predicting "low satisfaction".

Figure 1: Local explanations produced by LIME and SHAP methods illustrating model predictions of satisfaction classes for a particular counterfactual explanation instance, based on explanatory qualities.

overall behavior or general patterns, addressing broader questions such as "How does the model make decisions in general?". A global explanation could be a simplified surrogate model (such as a decision tree or rule set) that captures the overall decision-making logic of a complex model (R. M. Byrne 2023), or a visualization of structural components within the model (e.g., attention weights in neural networks). Local and global are not mutually exclusive. Often, multiple local explanations can be aggregated to identify global decision-making trends, as illustrated by the SHAP summary plot (Figure 2). For example, a study clustered numerous

local explanations (in the form of heatmaps) to uncover decision strategies within a vision model (Schwalbe et al. 2024). The choice between local and global explanations ultimately depends on user needs: local explanations support case-by-case accountability ("Why was my loan denied?"), while global explanations facilitate model validation, debugging, and identification of biases by revealing the overall decision rules and logic.

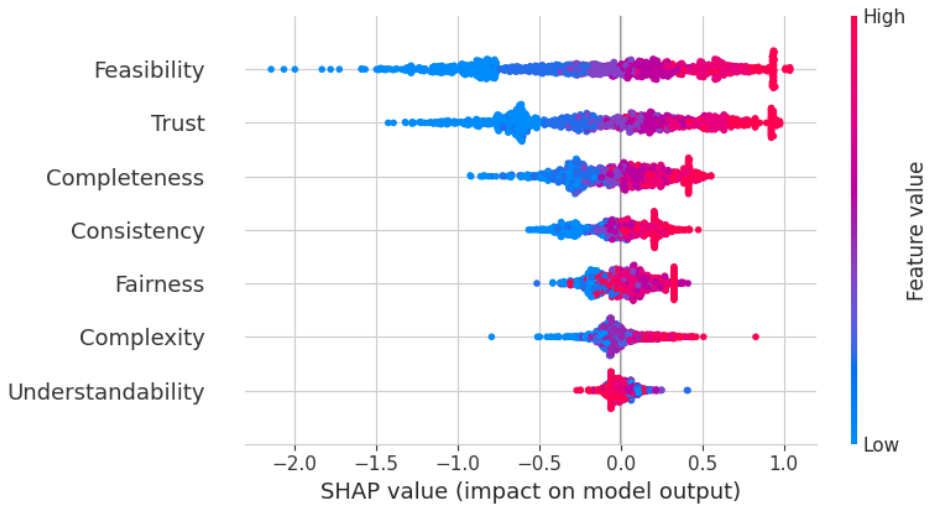


Figure 2: SHAP summary plot providing global insights by aggregating feature impacts across multiple predictions.

2.2.4. Model-agnostic and Model-specific explanations

Explanations can also be categorized based on the explainer's reliance on the internal structure of the model. **Model-agnostic** methods treat the target model as a black box, requiring only input-output interactions without assumptions about the model's internals. These methods typically employ perturbations, probes, or surrogate models to interpret the model's behavior. For example, LIME (Local Interpretable Model-agnostic Explanations) creates a local surrogate model, such as a linear model, around a specific input instance by querying the black-box model. Similarly, SHAP evaluates feature contributions based on Shapley values and can be applied universally across different model types. The primary advantage of model-agnostic approaches is their broad applicability. However, their potential downside is the risk of generating explanations that do not accurately reflect the true internal logic of the model (Rudin 2019), leading to possible misinterpretations and over-reliance (Ghassemi et al. 2021).

In contrast, **model-specific** methods (also known as white-box or decompositional explainers) are tailored to particular model types, leveraging internal parameters, gradients, or architectural details. An illustrative example is Grad-CAM (Gradient-weighted Class Activation Mapping), which is designed explicitly for

convolutional neural networks. Grad-CAM utilizes the CNN's internal feature maps and gradients to produce visual explanations of image predictions. While such methods provide valuable insights into the model's inner workings, they typically depend heavily on gradients or specific architectural properties. Recent developments in mechanistic interpretability, such as Anthropic's use of Sparse Autoencoders (SAEs), extend this concept and demonstrate a deeper form of model-specific explanation. SAEs decode internal activations of large neural networks into sparse, interpretable latent features, allowing researchers to identify and manipulate concept-level representations directly within the network. For instance, SAEs have enabled causal interventions to understand how specific internal features influence model outputs (Templeton et al. 2024).

2.2.5. Example-based and Feature-based Explanations

Explanations can also be categorized by what form of information they provide to the user. **Feature-importance-based** explanations highlight which input features most strongly influenced a prediction and, sometimes, quantify their contribution. Examples include attribution methods, such as saliency maps for images or feature importance scores for tabular data, as well as weights derived from surrogate models like LIME or SHAP (see Figure 2). These explanations typically address questions such as, "Which features drove the prediction?". For instance, a medical diagnosis model might clarify its decision by indicating that "symptom X and symptom Y were the strongest contributors."

In contrast, **example-based explanations** illustrate a model's reasoning through specific instances or examples. Instead of, or in addition to, presenting feature weights, the model might reference similar past cases, prototypes, or hypothetical scenarios. This category includes explanations like prototypes and exemplars (e.g., "This email was classified as spam because it resembles these known spam examples"), counterfactual examples (e.g., "If this input had been X instead of Y, the outcome would have changed"), and case-based reasoning that identifies influential training examples. Example-based explanations are intuitive, aligning closely with human reasoning, which often relies on analogy and examples. For instance, a vision model classifying an image as a "cat" might illustrate its decision by showing the training images of cats that activated similar internal neurons. Another practical scenario could involve a loan approval model providing a counterfactual explanation, such as, "If your income were 500 euros higher, the loan would have been approved." This approach offers a concrete, actionable insight rather than abstract feature importance alone. Modern surveys (e.g., Poché et al. 2023) emphasize that example-based explanations resonate naturally with users because humans inherently learn and understand concepts through examples.

2.3. Counterfactual explanations

This section introduces counterfactual explanations, a type of human-centric explanation that illustrates minimal changes required to alter a model's decision. It highlights their cognitive, philosophical, and practical significance, providing a historical overview as well as their formalization within Explainable AI.

2.3.1. The importance of Counterfactual Explanations

Counterfactual explanations describe how altering an input can change a model's output, answering the question "What should be the minimal change to the input required to achieve the desired outcome?". For example, "For this student to be accepted to the University of Tartu, the academic test score would need to be 85 instead of 65". Such explanations are contrastive (they explain why this outcome vs. an alternative) and selective (focusing on minimal feature changes), corresponding to (Miller 2019) desired properties of explanation and making them human-friendly. In the context of Explainable AI, counterfactual explanations are *example-based local explanations*, offering a concrete what-if scenario showing how to obtain a different result for a specific instance of interest rather than abstract feature importance measures.

Counterfactual explanations align with natural human reasoning, grounded in causal and philosophical traditions dating back to David Lewis's seminal work on counterfactual reasoning (Lewis 1973). These explanations help users to understand individual predictions by demonstrating small changes that would flip the model's decision, giving the explainee *actionable* recommendation to change their outcome. This aligns with how humans naturally reason about causation and explanations ("What if X had been different?") (Kshetry et al. 2024).

Cognitive and social science research suggests that effective explanations must be tailored to the explainee's context. Counterfactual explanations naturally fit these patterns by highlighting a minimal change that contrasts the factual outcome with a desired alternative. This resonates with cognitive psychology findings that counterfactual thinking (imagining "what if" scenarios) is a pervasive mode of human reasoning (R. M. Byrne 2016). People spontaneously construct alternative outcomes ("If only") to explain events and to learn from them, indicating that explanations grounded in such what-if scenarios will feel familiar and meaningful.

From a causality perspective, this counterfactual mode of reasoning is essential for understanding and explaining events. According to Judea Pearl's ladder of causality, correlations or feature importances alone rarely satisfy human curiosity because they do not answer these counterfactual "why" questions. Counterfactual explanations, by contrast, provide this causal and actionable insight by speaking the same "what if" language that our brains use when we seek explanations (Pearl 2009).

These explanations are particularly valuable in high-stakes domains, such as finance or employment decisions (Karimi et al. 2022), they serve as a form of

algorithmic recourse, advising those unfavorably treated by an AI what actions could lead to a positive result. They provide actionable insights to end-users and decision-makers. Moreover, counterfactual explanations support strategic planning for domain experts and decision-makers ("What should be the minimal change to meet the threshold?").

Additionally, by showing which feature changes significantly affect outcomes, counterfactual explanations help audit and *debug models*. For example, consistent recommendations to alter immutable or sensitive attributes (e.g., race or gender) may indicate underlying biases or fairness issues in the data or model training and are even argued to be a legal necessity under regulations (Sharma et al. 2019). Researchers have noted that counterfactual explanations enable a "guided audit" of the data and model behavior (Karimi et al. 2022), since they tie decisions to concrete conditions. Moreover, Guidotti (2022) highlights that the lack of transparency in opaque models is a practical and ethical issue, and counterfactual explanations address this in a human-interpretable way.

While the concept of counterfactual reasoning has deep roots in philosophy and causality, counterfactual explanations as a specific XAI method gained attention thanks to the work of Wachter et al. (2017), who formally introduced "counterfactual explanations" and argued for their use under the right to explanation in GDPR. Further momentum was gained thanks to Miller (2019), who highlighted the human nature of such a method and emphasized the cognitive alignment of counterfactual explanations with human reasoning. Multiple influential papers refined the idea, such as Russell (2019), which proposed search algorithms for diverse and coherent counterfactuals using mixed-integer programming. Mothilal et al. (2020) introduced DiCE to generate a diverse set of plausible counterfactual explanations and provided a well-implemented library that quickly gained widespread use. Following these works, counterfactual explanations became a core topic in explainable AI research. Comprehensive surveys (Artelt et al. 2019; Guidotti 2022; Verma et al. 2020) classified counterfactual examples as a fundamental approach to explain AI decisions. Depending on the level of access to the model, counterfactual explanation algorithms are developed as both model-agnostic and model-specific. Over just a few years, counterfactual explanation has evolved from a philosophical notion to a practical toolkit for interpretable machine learning, with dedicated surveys consolidating terminology and methods. Today, hundreds of algorithms are developed that use the concepts of causality, fairness, and user-centric explainability.

2.3.2. Formalization of Counterfactual Explanations for Tabular Data

The formalization of counterfactual search involves framing it as an optimization problem aimed at identifying an input instance minimally different from the original instance that achieves a targeted prediction outcome. Formally, let f represent a trained predictor function that maps the input space $\mathcal{X}$ to the out-

put space $\mathcal{Y}$, i.e., $f : \mathcal{X} \to \mathcal{Y}$. Given an original input instance (a factual point) $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^n$, the objective is to identify a counterfactual point $x^* = (x_1^*, x_2^*, \ldots, x_n^*) \in \mathcal{X} \subseteq \mathbb{R}^n$ that predicts the desired label y while minimizing a loss function and a distance penalty:

$$\arg\min_{x^*} \Big(\text{loss}(f(x^*), y) + \lambda \text{dist}(x^*, x) \Big), \tag{2.1}$$

where:

- $\text{loss}(f(x^*), y)$ is the loss term that penalizes deviation from the desired outcome y.
- $\text{dist}(x^*, x)$ quantifies the distance between the original point x and the counterfactual point x^*.

The hyperparameter λ balances the trade-off between achieving the desired prediction and remaining close to the original input. In practice, additional constraints can enforce that $f(x')$ explicitly belongs to the target class.

The choice of the **loss** function depends on the task. Classification tasks commonly employ cross-entropy or probability-shifting constraints toward the target class (Molnar 2020, Chapter 15). For regression tasks, the loss often measures proximity to a desired numerical target (between $f(x^*)$ and y using Mean Square Error (MSE) loss, for instance). Sometimes, the regression task is divided into value ranges of outcomes, and counterfactual explanation search can be performed similarly to classification, treating flipping the outcome to the desired prediction range.

The choice of **distance** $\text{dist}(x^*, x)$ is crucial to quantify similarity. Standard options include L_2 (Euclidean distance), L_1 (Manhattan distance) norms on feature differences, sometimes normalized by feature variability to make units comparable. For example, Wachter et al. (2017) used a weighted Manhattan distance scaled by the median absolute deviation. In the case of mixed feature types (numerical and categorical), there is a natural division between feature distances. Sometimes Gower or Hamming distances are applied. But more commonly, an L_0 norm is used to detect category changes:

$$dist_{cat}(x, x^*) = \frac{1}{m} \sum_{j=1}^{m} I(x_j^* \neq x_j) \tag{2.2}$$

Categorical variables require special consideration during optimization due to their discrete nature. For example, the feature 'Job' taking values such as 'Teacher' or 'Engineer' cannot be smoothly transitioned through gradient-based steps. Two common approaches are: (1) One-hot encoding categories and treating the 0/1 entries like continuous variables during optimization (with extra constraints to keep one-hot validity), or (2) using mixed-integer optimization where binary decision variables represent categorical flips. The distance for a categorical feature is usually defined as 0 if the category is unchanged or 1 if it is different (effectively an

L_0 or Hamming distance on that feature). This can be combined with numeric distances for continuous features in a single objective (Molnar 2020, Chapter 15).

Given the crucial nature of the distance between original point and counterfactual distance, multiple advanced methods were proposed accounting for data manifold similarity, such as Mahalanobis distance (Kanamori et al. 2020), autoencoder-based distances (Dhurandhar et al. 2018), Dijkstra (Poyiadzi et al. 2020), and others. In Chapter III, we discuss such advances more in depth, as our contribution relates to employing different measures of proximity.

To obtain intuitive explanations, additional regularizers or constraints are introduced: **Sparsity** can be enforced by penalizing the L_0 norm (or by incrementally searching for solutions with one feature change, two feature changes, etc.). **Coherence** or **Plausibility** can be enforced by constraining x^* to lie within realistic bounds. That implies no negative values for inherently positive features, no changes to immutable attributes like race or gender, etc. One may also add a penalty if x' has low likelihood under the training data distribution (Molnar 2020, Chapter 15).

To handle potential model changes or data shifts, **robust** recourse methods were proposed that optimize for worst-case scenarios or verify validity under model perturbations. For example, Upadhyay et al. (2021) solves a min–max problem to find counterfactuals that remain valid even if the classifier is retrained with slightly different parameters. Some works enforce Δ-robustness by requiring a counterfactual to hold across the entire family of perturbed models (Jiang et al. 2024). Additionally, Artelt et al. (2021) studied the robustness of counterfactual explanations, highlighting how minor input perturbations can cause significant or arbitrary changes in counterfactual recommendations, affecting individual fairness. Some approaches incorporate uncertainty estimates, in particular, using probabilistic models or Bayesian neural networks (Antorán et al. 2020) to avoid counterfactuals in regions where the model is not confident, thus focusing on more robust explanations.

Finally, providing users with multiple actionable alternatives highlights the importance of **diversity** among counterfactual explanations. Diversity ensures users receive distinct recourse options, enabling them to choose the most feasible or preferable action. For instance, Mothilal et al. (2020) a diversity-promoting regularizer based on determinantal point processes, explicitly designed to generate diverse sets of counterfactuals. Interestingly, such diversity can itself boost robustness, as in Leofante et al. (2024) it was proved that while a single counterfactual may fail under slight model or input drifts, a sufficiently diverse counterfactual set can guarantee that at least one valid explanation persists for similar inputs.

2.3.3. Counterfactual explanations search methods

After defining the optimization objective, several algorithmic approaches can be applied to navigate counterfactual search.

Gradient-Based Optimization can be applied if the model f is differentiable or can be approximated with a different surrogate. It is common to use gradient-based methods to minimize the counterfactual loss if all terms are differentiable. This approach was taken by Wachter et al. (2017), who solved counterfactual objective function of the form

$$\min_{\mathbf{x}'} \lambda (f(x') - y')^2 + d(x, x')$$

with gradient descent, directly optimizing prediction proximity and distance terms.

Evolutionary and Heuristic Search can handle non-differentiable objectives or incorporate multiple objectives used by many approaches. For example, Dandl et al. (2024) proposed a multi-objective genetic algorithm explicitly optimizing four criteria: validity of outcome, proximity, sparsity, and Plausibility. In this method, candidate solutions x' evolve through iterative mutations of feature values, retaining solutions based on their fitness relative to these criteria. This evolutionary approach naturally supports generating a diverse set of explanations in a single run, offering users multiple actionable options. However, this flexibility comes at an increased computational cost, as heuristic algorithms typically require significant computational resources to converge on global optima reliably. Aside from genetic algorithms, other heuristics like simulated annealing or beam search have also been used to prioritize different biases (some focusing on robust solutions that remain valid under slight model changes, others generating a set of diverse explanations so the user can choose the most actionable one).

Mixed-Integer Programming (MIP) is another class of methods that allows the use of solvers to guarantee optimality. These methods are instrumental when the model is interpretable or constrained enough to encode (e.g., decision trees, linear models, or rule-based classifiers), or when we can linearize the decision boundary conditions. For instance, Russell 2019 used MIP to find coherent counterfactual explanations that respect data constraints efficiently. By introducing binary variables for whether a feature is changed, one can directly minimize the L_0 count of changes (ensuring sparsity) and at the same time enforce that $f(x')$ meets the desired outcome exactly. MIP-based approaches easily handle categorical features and can incorporate logical constraints (like "education level cannot decrease" or "if feature A changes, feature B must also change accordingly"). They tend to produce the truly minimal counterfactual (in terms of a chosen cost) due to the exhaustive nature of the solver search, and can ensure feasibility by encoding domain knowledge. The downside is that exact solvers may not scale well to very complex models or high-dimensional feature spaces (the problem can become NP-hard). Still, for moderately sized problems, they provide strong guarantees.

Sampling-Based Methods, such as the Growing Spheres algorithm proposed by Laugel et al. (2018), represent an alternative approach. Although not explicitly using the term "counterfactual" in their original work, their methodology aligns closely with counterfactual search objectives. Instead of direct optimization,

Growing Spheres define a proximity-based loss that favors counterfactuals with minimal feature modifications. This method begins by drawing a sphere around the point of interest and sampling within this sphere. If no suitable counterfactual is found, the sphere is iteratively expanded or contracted accordingly until a sparse counterfactual is discovered. This approach identifies sparse counterfactuals without explicitly relying on gradient information, aiding methods where differentiability is limited.

2.3.4. Counterfactual Explanations for Image Data

While counterfactual explanations have been mainly explored in the context of tabular data, recent years have witnessed increasing efforts to adapt these methods to high-dimensional domains such as images (Akula et al. 2022; Augustin et al. 2022; Vandenhende et al. 2022). In this setting, a counterfactual explanation for an input image $x \in \mathbb{R}^{H \times W \times C}$ of height H, width W, and C channels takes the form of a perturbed version x^* that leads a predictive model f to change its output to a desired class y, while remaining similar to the original image. Unlike tabular data, there is no commonly accepted formalization for counterfactual generation in images. However, most approaches agree conceptually on two fundamental components: a classification loss, ensuring that the generated counterfactual achieves the desired target prediction, and a distance measure, enforcing visual similarity to the original image. For instance, OCTET (Zemni et al. 2023) explicitly formulates these two objectives using a decision loss (L_{decision}) to drive the model towards the desired prediction and a latent-space distance term (L_{dist}) to maintain similarity to the original image representation. However, having only loss and distance measures leads to adversarial-like solutions that technically satisfy the optimization, but do not give a meaningful explanation. To address these shortcomings, recent methods incorporate additional constraints and regularizers. For example, Vandenhende et al. (2022) introduced semantic consistency losses to ensure modifications are semantically meaningful rather than arbitrary pixel-level perturbations. Similarly, Charachon et al. (2022) and Singla et al. (2023) proposed a cycle-consistency loss to preserve specific original image features, ensuring counterfactual transformations are minimal and reversible, enhancing interpretability and realism, particularly relevant in medical contexts.

Therefore, let's define a counterfactual objective function with three main terms. The goal is to find a modified image x^* such that:

$$\arg\min_{x^*} \quad \mathcal{L}\text{cf}(x^*, y^*) + \lambda \cdot \mathcal{D}(x^*, x) + \beta \cdot \mathcal{R}(x^*), \tag{2.3}$$

where:

- $\mathcal{L}\text{cf}(x^*, y^*)$ is the loss function encouraging $f(x^*) = y^*$, typically a cross-entropy or classification loss.
- $\mathcal{D}(x^*, x)$ is a similarity loss penalizing visual dissimilarity from the original image.

- $\mathcal{R}(x^*)$ is a regularization term to enforce realism or prior knowledge (e.g., anatomical constraints).

Optimization: Given the importance of realism, generative models such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models have become central to image counterfactual research. VAEs encode images into latent representations and seek nearby latent codes that decode into realistic counterfactual images. VAEs have the advantage of disentangling factors of variation to some extent, so that specific latent directions can correspond to human-interpretable changes. However, optimizing solely in latent spaces without careful regularization can lead to unrealistic solutions outside the learned data distribution.

GAN-based counterfactual methods take advantage of GANs' ability to generate high-fidelity images. A common approach involves finding latent codes corresponding to the input images, which are then modified to achieve the desired predictions (Nemirovsky et al. 2022). In particular, conditional GAN methods (cGAN) allow for generating localized modifications through learned masks or residual generators (Chang et al. 2018). Charachon et al. (2022) and Singla et al. (2023) use cGAN models with a cycle-consistent loss function to introduce reversibility constraints to ensure minimal and meaningful edits, preserving class-related features while maintaining realism.

Advanced GAN architectures incorporate semantic constraints by explicitly segmenting images into semantic or object-centric components, altering only the relevant parts. Methods such as OCTET use object-centric latent representations to ensure sparsity in changes (Zemni et al. 2023). Similarly, approaches like STEEX use segmentation-to-image GANs to modify appearance features in specific regions, preserving overall image structure (Van Looveren et al. 2021b).

Diffusion models are a recent generative approach explored for counterfactual explanations. These models iteratively denoise images, guided by classifier gradients to subtly push generated images toward desired predictions. For instance, Jeanneret et al. (2022) introduced DiME, using classifier guidance during reverse diffusion to gradually approach the decision boundary. Diffusion models have stable training dynamics and naturally allow diverse outputs due to their stochastic sampling processes. They also support flexible conditioning, such as semantic or textual guidance, enabling targeted counterfactual edits (e.g., medically meaningful changes).

Another direction emphasizes explicitly interpretable semantic concepts or attributes, rather than pixel-level modifications alone. Conceptual Counterfactual Explanations (CCE), for instance, employ Concept Activation Vectors (CAVs) to adjust image attributes that align closely with human understanding, such as object textures or anatomical features (Akula et al. 2022). Such concept-based approaches are particularly advantageous in domains where interpretability at the conceptual level significantly impacts practical decision-making. For example, to

explain why a model classifies an image as a "sparrow" rather than a "robin", their approach might find the head region of the bird in the image and replace it with the head of a robin from another image; the modified image is then classified as a robin, showing that the head color was the discriminative feature. This case-based technique yields a counterfactual that is itself a composite of real image parts, maintaining realism and clearly pointing to the region that caused the change. More generally, concept-based counterfactuals leverage interpretable features such as "has stripes", "wearing glasses", or "dark background".

2.4. Evaluation of Counterfactual Explanations

This section examines the main challenges in evaluating counterfactual explanations and surveys existing methods used for their assessment.

2.4.1. Common quantitative metrics

Counterfactual explanations are typically evaluated using several quantitative criteria. These criteria are based on assumptions that a good counterfactual should successfully flip the model's prediction (validity); require only minimal change from the original instance (proximity and sparsity; correspond to a plausible and realistic data point (Plausibility, constraint violation); if multiple explanations are provided, offer diverse alternatives (diversity) (Guidotti 2022). Less frequently considered, robustness reflecting the stability of explanations under small perturbations, and fairness of explanations capturing whether explanations consistently treat different groups equitably (Goethals et al. 2024; Jiang et al. 2024). We review each of these metrics below.

By definition, a counterfactual explanation must achieve the desired outcome. For instance, it should change the model's prediction from a negative to a positive decision. **Validity** (sometimes called success rate) checks whether the generated counterfactual is indeed classified in the target class by the model (Guidotti 2022; Karimi et al. 2022; Stepin et al. 2021; Verma et al. 2020).

$$\text{Validity} = \frac{\text{Number of successful counterfactuals}}{\text{Total number of counterfactuals generated}} * 100\% \tag{2.4}$$

Most counterfactual methods guarantee validity by construction, but in evaluation, one may measure the fraction of test instances for which at least one valid counterfactual can be found. This is often referred to as **coverage**. A higher coverage or validity ratio indicates better performance.

Proximity similarly arises from the definition of counterfactual explanation. Proximity measures the magnitude of the overall change between the original instance and the counterfactual. Intuitively, the counterfactual should be as close as possible to the original case to minimize the cost or effort of change. However, the selection of distance measures for evaluation often appears arbitrary. The weighted L_1 norm is widely used for continuous features (Wachter et al. 2017), together with

the L_0 norm for categorical features (Guo et al. 2023; Mahajan et al. 2019; Mothilal et al. 2020). Sometimes, multiple distances, such as Euclidean, Cosine, are used in parallel with L_1 for evaluation (Lucic et al. 2022), or diffusion distance is used together with L_1 (**Domnich** et al. 2024). Alternatively, Mahalanobis distance is used (Kanamori et al. 2020) for both optimization and evaluation.

The most common **The Weighted L1** is defined by adjusting the L1 norm with the inverse of the Median Absolute Deviation (MAD) for each feature to penalize large deviations. The formula is given by:

$$L1_{\text{Wachter}}(x, x') = \sum_{i=1}^{M} \left(\frac{|x_i - x'_i|}{\text{MAD}_i} \right) \tag{2.5}$$

where M is total number of points, MAD_i is the median absolute deviation of the i-th feature across the dataset, and x_i and x'_i are the values of the i-th feature in the original and counterfactual instances, respectively.

L0 Categorical counts the number of altered features and is defined as:

$$L0(x, x') = \|\{i \mid x_i \neq x'_i\}\|_0 \tag{2.6}$$

indicating the count of non-zero differences between corresponding features of x and x'.

While proximity deals with the amount of change, **sparsity** focuses on the number of features changed. A counterfactual is sparse if it alters only a few features of the original instance, ideally the minimal subset needed to achieve the desired outcome (Verma et al. 2020; You et al. 2023).

Not all mathematically valid counterfactuals make sense in reality. **Plausibility** measures how realistic or "data-like" a counterfactual is, given what is known about the domain. A counterfactual is considered plausible if its feature values represent a coherent, likely scenario (i.e., it lies on the data manifold of real instances). One way to operationalize Plausibility is to measure the distance from the counterfactual to the nearest data point in the training set, increasing the confidence that it is realistic (Guidotti 2022). Sometimes different names are used for this metric, such as **Coherence**, or connectedness (Rasouli et al. 2024), which also refers to whether a counterfactual point belongs to the desired class or data distribution. Some methods use density estimates or generative models. For instance, computing the likelihood of the counterfactual under a generative model of the data to create Interpretability Metrics (IM1 and IM2) proposed in (Van Looveren et al. 2021a), or applying outlier detection methods like Local Outlier Factor (Kanamori et al. 2020). Directional coherence requires alignment with the marginal direction of increased label likelihood (**Domnich** et al. 2024). In contrast, (Mahajan et al. 2019) opts for causal constraint satisfaction, focusing on the causal Plausibility of counterfactuals rather than their statistical alignment. Clearly, there is no standardized approach to evaluating Plausibility in the field.

Actionability evaluates whether the suggested feature changes in a counterfactual are feasible for an individual to carry out in practice. Even a plausible counterfactual might recommend changes to immutable or non-controllable attributes (e.g., "change your age from 50 to 40"). To account for actionability, methods typically define a set of actionable features or constraints (often based on domain knowledge), indicating which attributes can be changed and in what direction or range, and actionability is evaluated as constraints violation count (Mahajan et al. 2019).

Robustness, a less commonly used metric, assesses whether outcomes are consistent under conditions of noise or following model retraining. This includes evaluating counterfactual stability, Validity after Retraining (VaR), or justification score (Guo et al. 2023; Jiang et al. 2024; Laugel et al. 2019b). There are also specialized metrics, such as Value at Risk, that address specific risks (Wu et al. 2024).

Many counterfactual explanation methods can return not just one but several alternative counterfactuals for a given instance. For such cases, it is important to evaluate diversity. **Diversity** measures the extent to which these multiple explanations differ from one another. To quantify diversity, researchers often compute pairwise distances between all counterfactuals in the set and take an average (the larger this average distance, the more diverse the explanations) (Ley et al. 2022; Mothilal et al. 2020; Rasouli et al. 2024).

2.4.2. Qualitative user studies

While quantitative metrics are commonly employed to assess counterfactual explanations, incorporating qualitative user studies is crucial for capturing subjective human preferences and perceptions (Keane et al. 2021a; Kirsch 2017; Longo et al. 2024). However, surprisingly few studies have included user evaluations. A survey of counterfactual methods Keane et al. (2021a) found that only 21% out of 100 surveyed studies had performed user tests of the specific counterfactual algorithm. Furthermore, many of those studies test the use of counterfactual explanations vs no-explanations rather than comparing different methods, leaving only 7% of papers that report user evaluations for benchmarking between different counterfactual algorithms.

Recent user studies, primarily conducted on tabular data, provided insights into user perceptions and the effectiveness of counterfactual explanations. For example, Warren et al. (2022) compared counterfactual and causal explanations in a study with 127 participants, finding that counterfactual explanations enhanced both prediction accuracy and subjective satisfaction and trust, particularly for categorical features. Similarly, Bove et al. (2023) investigated interface design with multiple counterfactual examples, demonstrating that presenting plural explanations significantly improved user understanding and satisfaction in a lab study involving 112 participants. Förster et al. (2021) assessed the coherence of counterfactual ex-

planations with 46 participants, concluding that coherent explanations effectively explained realistic scenarios. Spreitzer et al. (2022) evaluated practicality perceptions of two counterfactual explanation methods with 135 participants, noting variations in perceptions depending on the classification tasks involved. Ghazimatin et al. (2020) further examined the practicality of explanation through an online study involving 500 participants, highlighting the importance of minimal and actionable counterfactuals.

In non-tabular domains such as image and multimedia tasks, user studies have also provided valuable insights. Akula et al. (2022) benchmarked different counterfactual algorithms using image data, measuring justified trust and explanation satisfaction. Goyal et al. (2019) introduced visual counterfactuals for image-based Question Answering (QA) and found that these counterfactual images improved participants' task accuracy compared to no explanations.

Textual and other domains have seen fewer qualitative evaluations that specifically target counterfactual explanations. One of the exceptions is Tešić et al. (2022), which investigated the cognitive impacts of counterfactual explanations on participants' causal beliefs in AI predictions through multiple controlled experiments. This research highlighted how counterfactual explanations might influence user beliefs, underscoring the importance of studying cognitive and psychological effects beyond only task performance.

Generally, user studies evaluating AI explanations often emphasize limited dimensions, typically user satisfaction and trust in the system that generates the explanations (Mueller et al. 2019). Such a narrow focus often overlooks broader explanatory qualities, relying on assumptions about human preferences that may not always reflect genuine user satisfaction. Psychological research emphasizes the context-dependent nature of human preferences, influenced by factors such as presentation style and cognitive biases, especially when users lack well-defined preferences (Covell 2019; Tversky et al. 1993). Consequently, user studies that do not account for explanatory dimensions in a comprehensive way risk providing an incomplete understanding of human judgment. Thus, significant gaps remain in identifying which explanatory features are most crucial to users.

3. ENHANCING COUNTERFACTUAL ALGORITHM (PUBLICATION I)

3.1. Introduction

This chapter introduces the first contribution of this thesis, which addresses **RQ1**: Can we extract more feasible and coherent explanations compared to existing methods? Specifically, we enhance a counterfactual explanation algorithm by integrating human cognitive preferences. We focus on two shortcomings in existing techniques: feasibility (Poyiadzi et al. 2020) and coherence with human reasoning (Keil 2006; Miller 2019).

To tackle feasibility, we rethink the concept of proximity by redefining what it means for a counterfactual point to be "close" in a realistic sense. We achieve this by employing diffusion distance, which considers feasible transitions within the data manifold. Furthermore, to improve coherence with human reasoning, we introduce directional coherence, which ensures that suggested counterfactual changes align with realistic feature-outcome relationships and thus better reflect real-world causal constraints.

3.2. Motivation

Counterfactual explanations are essential tools in XAI for explaining model decisions by answering the question "What should be the minimum input change in order to flip the decision outcome?". Central to the concept and definition of counterfactual explanations is the notion of distance, which determines how minimal or realistic such changes are. Most counterfactual algorithms rely on either weighted L1 (Mothilal et al. 2020; Wachter et al. 2017), or L2 distance (Parmentier et al. 2021), or their Elastic net combination of L1, L2 (Van Looveren et al. 2021a), or in fewer cases, Mahalanobis distance (Kanamori et al. 2020), often without explicitly considering how well these distances reflect the underlying data structure and intended counterfactual usage. On the latter point, different types of use cases for counterfactual explanations include: 1) debugging models to identify whether spurious correlations influence the decision boundary or whether the model relies on biased features; 2) supporting actionability by proposing realistic changes that the explainee can make. For the second purpose, counterfactuals must be reachable from the original instance via feasible paths in the data manifold, suggesting realistic transitions and trajectories. The FACE method (Poyiadzi et al. 2020) introduced feasibility constraints using k-NN graphs and Dijkstra's algorithm, ensuring counterfactuals follow paths through actual data points. However, FACE's reliance on existing data points and direct shortest paths limits its applicability, especially in high-dimensional or sparse datasets. Traditional approaches often generate counterfactuals in low-density, unrealistic regions, failing to reflect genuine actionable possibilities. Motivated by this gap identified in RQ1,

we employ **diffusion distance**, which enhances feasibility by identifying shortest transitions through multiple realistic paths within the data distribution. Unlike Euclidean metric, diffusion distance prioritizes transitions between data points that are interconnected by many short paths, bringing closely connected points nearer in distance. This ensures that counterfactual suggestions lie within denser, more realistic regions that accurately reflect the underlying data manifold.

Additionally, human reasoning inherently expects explanations to be both internally and externally coherent (Zemla et al. 2017). Cognitive science research emphasizes that effective explanations follow intuitive principles of causality and consistency (Rasouli et al. 2024). In practice, this means the reasoning in an explanation should not conflict with itself or with general causal knowledge. We observed that most methods, if they consider coherence, model it primarily in terms of feature distributions, often overlooking coherence concerning the label or outcome itself. For example, people find it counterintuitive if a loan approval explanation suggests increasing the applicant's income and simultaneously decreasing a co-applicant's income. It sounds conflicting, since decreasing co-applicant's income does not "fit the causal mechanisms we expect, connecting to a goal or purpose we understand" (Keil 2006). While such a data point exists in the data, it represents an undesirable direction for an explanation if actionability is prioritized. Keil emphasizes that it is enough for an explanation to be considered good simply because it is faithful to the model, because it should help the user to understand the decision. Inspired by this intuition of fitting counterfactual explanation to the goal and not contradicting known positive influences, we propose the notion of **directional coherence**, to ensure counterfactuals align with expected marginal feature effects on predicting the desired class. Therefore, addressing the coherence aspect of **RQ1**, specific feature changes should consistently drive the prediction in a predictable direction, and counterfactuals should respect these intuitions.

3.3. Main findings

Through empirical evaluation, we demonstrate the potential of integrating diffusion distance and directional coherence into the counterfactual explanation search, resulting in the creation of the CoDiCE (Coherent Directional Counterfactual Explainer) framework. We benchmarked CoDiCE against several established methods (DiCE, FACE, Prototypes, Growing Spheres) across multiple datasets (Diabetes, Breast Cancer, Adult Income, German Credit, COMPAS, and Energy Consumption) for both classification and regression tasks. Detailed findings are presented in the following sections.

3.3.1. Feasibility with Diffusion Distance

Diffusion distance is introduced as a means to ensure counterfactuals are plausible and lie on a feasible path within the data manifold. Unlike simple Euclidean distance, diffusion distance considers the connectivity of data points, effectively ac-

counting for multiple short hops through high-density regions instead of a straight-line jump through space.

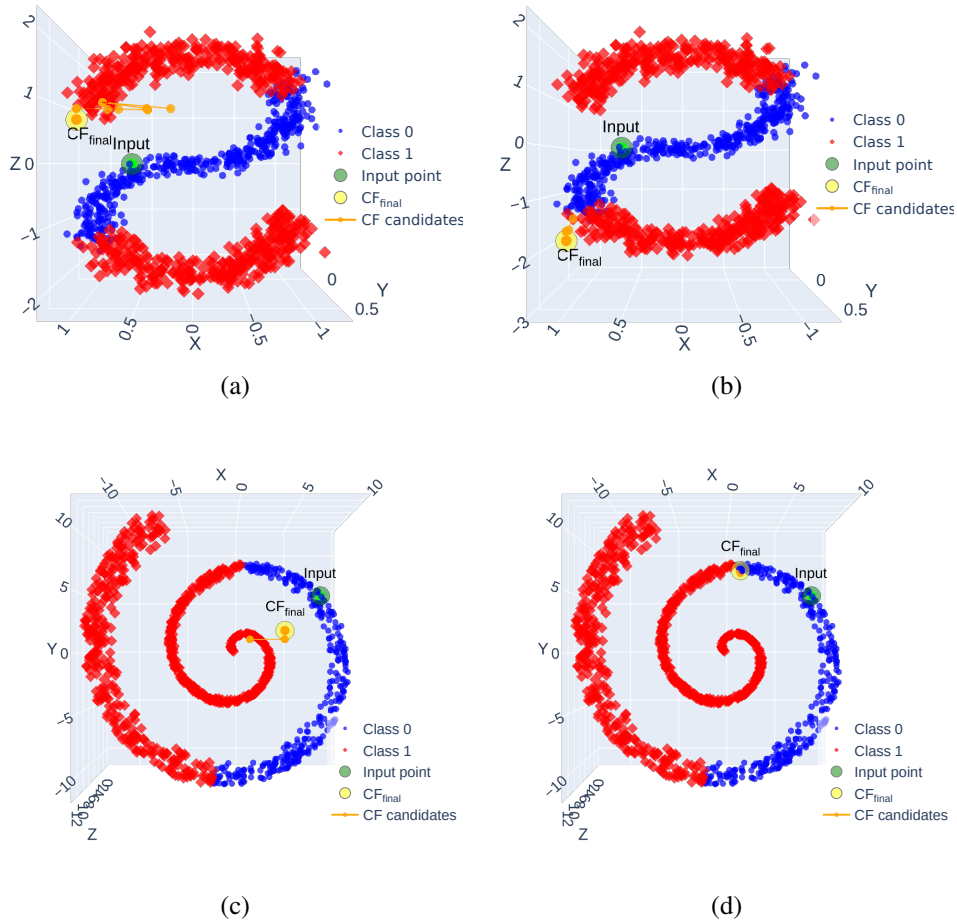


Figure 3: Counterfactual search on synthetic datasets with L_1 (left: a, c) and diffusion distance (right: b, d) (**Domnich** et al. 2024).

Formally, diffusion distance can be defined between two points x and y at time t as:

$$D_{\text{diff}}(x,y,t)^2 = \sum_z \frac{(p_t(x|z) - p_t(y|z))^2}{\phi_0(z)}, \tag{3.1}$$

where $p_t(x|z)$ represents the probability of transitioning from point z to x in t steps following a diffusion process (random walk on the graph), and $\phi_0(z)$ is the stationary distribution of the diffusion process at point z. We used diffusion distance with Self-Tuning Kernel, which is used for local scaling to adjust dynamically to the variance in the data.

Figure 3 provides a visualization: using diffusion distance for counterfactual

search avoids traversing low-density (sparse) regions that would lead to unrealistic instances for synthetic datasets (S-shape and swiss roll). Instead, it finds pathways through the dataset that connect the original instance to a plausible counterfactual. This approach is akin to exploring a map of the data: even if two points are not directly neighbors, there may exist a series of small steps (intermediate data points) linking them. Diffusion distance leverages this by bringing closely connected points nearer in the distance metric, meaning a counterfactual is considered "closer" if you can travel to it via many short jumps through actual data points. In practical terms, a counterfactual generated with this method respects the underlying data distribution – for example, in a health dataset, suggesting a gradual change (through realistic intermediate health states) rather than an abrupt, implausible leap. By using diffusion distance, the counterfactual search identifies transitions that are more feasible and actionable in reality. This results in counterfactual explanations that make sense from the perspective of data geometry and represent the realistic scenarios a person could take.

3.3.2. Formal Definition of CoDiCE counterfactual algorithm

We denote by f a trained predictor function that maps the input space $\mathcal{X}$ to the output space $\mathcal{Y}$, i.e., $f : \mathcal{X} \to \mathcal{Y}$. Given a factual point or the original input point $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^n$, our objective is to identify a counterfactual point $x^* = (x_1^*, x_2^*, \ldots, x_n^*) \in \mathcal{X} \subseteq \mathbb{R}^n$ that yields the desired label y while minimising a weighted sum of diffusion distance, sparsity, and directed coherence penalties. The optimization problem is defined as follows:

$$c = \arg\min_{x^*} \Big(\text{loss}(f(x^*), y) + \lambda_1 \text{diffusion\_dist}_{cont}(x^*, x) + \lambda_1 dist_{cat}(x^*, x) + \lambda_2 \text{sparsity}(x^*, x) + \lambda_3 (1 - \text{dcoherence}(x^*, x)) \Big) \tag{3.2}$$

where:

- $\text{loss}(f(x^*), y)$ is the loss term that checks if the counterfactual outcome is equal to the desired outcome. We utilize commonly used loss measures, hinge-loss for classification and mean squared error for regression.
- $\text{diffusion\_dist}(x^*, x)$ quantifies the diffusion distance between the original point x and the counterfactual point x^*.
- $\text{dist}_{cat}(x^*, x)$ quantifies the distance between categorical features with l_0 norm weighted by number of categories.
- $\text{sparsity}(x^*, x)$ computes the l_0 distance to count the number of features that have been modified.
- $\text{dcoherence}(x^*, x)$ assesses the directional coherence by aligning the joint direction of the counterfactual point with its marginals. Since we are interested in minimizing the objective function, we take the penalty measure $(1 - \text{dcoherence})$.

The terms are weighted by hyperparameters λ_1, λ_2, and λ_3, which can be adjusted or set to 0 if a particular constraint is not applicable.

Mathematically, we formulate directional coherence as a term that quantifies the preference for alignment between joint and marginal directional changes in the feature space necessary to achieve a counterfactual outcome. For clarity, we introduce here the case of a classifier. The corresponding formulation for a regression model is a straightforward extension.

As for directional coherence, the goal is to evaluate the coherence of the transition from x to x^* in achieving a specified outcome label y with a set of expected marginal transitions:

$$\{x_i \rightarrow x_i' \mid f(y|x_1, x_2, \ldots, x_i', \ldots, x_n) \geq f(y|x_1, x_2, \ldots, x_i, \ldots, x_n), 1 \leq i \leq n\}.$$

Although marginal transitions are typically derived from the model, user-specified marginal transitions, when provided, will take precedence over those suggested by the model.

Then, the Directional Coherence score counts the number of features which have aligned marginal ($'$) and joint ($*$) directions to increase the model's prediction probability towards the desired outcome y:

$$dcoherence = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}\left(\left(x_i^* - x_i\right)\left(x_i' - x_i\right)\right). \tag{3.3}$$

3.3.3. Directional Coherence

Directional coherence ensures that recommended feature changes in a counterfactual align with intuitive and learned relationships. It helps to avoid contradictions with known patterns of individual feature influence on the prediction.

Figure 4 illustrates this concept: an input point (Class 1) has two possible counterfactuals (CF1 and CF2) that are at equal distance from it, aiming for Class 2. The data trend indicates that increasing Feature 1 and Feature 2 correlates with a higher probability of Class 2. Consequently, CF2 (which increases both features) is directionally coherent, while CF1 (which decreases Feature 1) is not, as CF1's change goes against the expected effect of Feature 1. Enforcing directional coherence means the counterfactual search favors solutions like CF2, where the joint changes in features agree with their individual positive contributions to the outcome. This leads to explanations that make intuitive sense (e.g., "increase income to improve loan approval odds" rather than suggesting a decrease) and avoids counterintuitive recommendations. By aligning counterfactual changes with the model's marginal feature effects, the explanations remain consistent with human expectations of causality (with respect to output) and the monotonic influence of a feature. Essentially, the directional coherence bias steers the explanation towards what a human would consider an intuitive direction of change for achieving the desired outcome.

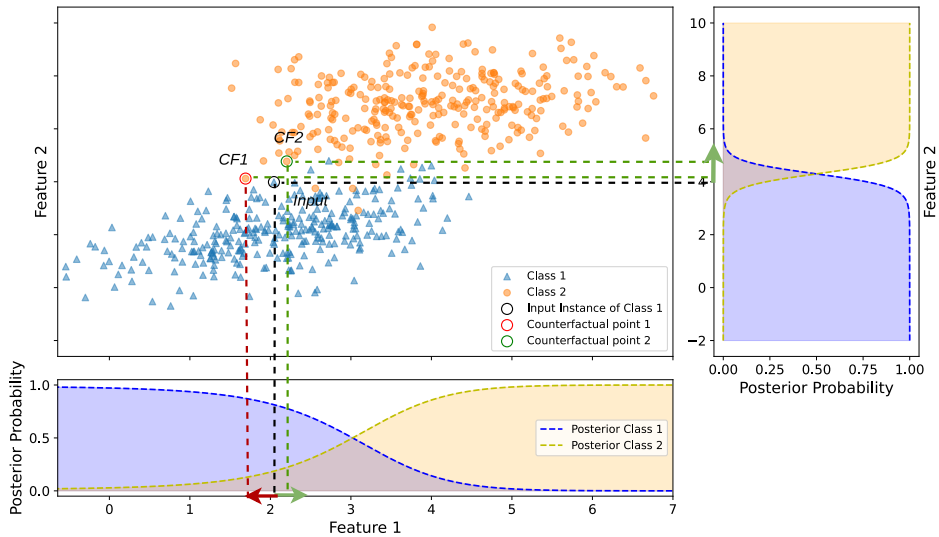


Figure 4: Illustration of Directional Coherence in Counterfactual Analysis. For an input point in Class 1, two counterfactual points CF_1 and CF_2 are at equal distances. CF_1 is deemed incoherent as it suggests decreasing Feature 1, contrary to its expected effect on increasing the probability of Class 2. While CF_2 aligns the direction of feature changes with the joint effect, resulting in a coherent counterfactual (**Domnich** et al. 2024).

3.3.4. Benchmarking Against Existing Methods

We benchmarked our method, CoDiCE, against several well-established counterfactual explanation algorithms, including DiCE (Mothilal et al. 2020), FACE (Poyiadzi et al. 2020), Guided Prototypes (Van Looveren et al. 2021a), and Growing Spheres (Laugel et al. 2018). The benchmarking was conducted across widely-used datasets in counterfactual explanation literature (Pawelczyk et al. 2021; Rasouli et al. 2024) representing diverse real-world contexts: continuous-feature classification tasks (Diabetes, Breast Cancer; Table 1), mixed-type classification tasks (Adult Income, German Credit, COMPAS; Table 2), and rarely covered in literature regression task (Energy Consumption; Table 3). Diabetes and Breast Cancer datasets were chosen due to their continuous features and clear interpretability of feature changes in healthcare applications. Adult Income, German Credit, and COMPAS datasets represent mixed-type feature challenges, reflecting common scenarios in fairness-sensitive applications. The Energy Consumption dataset provided a regression setting representative of real-world application scenarios where gradual actionable changes are crucial, further validating CoDiCE's applicability across varied real-world problems.

We assessed the algorithms using standard metrics, such as validity and proximity (distance measures), for comparability with established methods. Moreover, we added our novel metrics: diffusion distance and directional coherence to evalu-

ate feasibility and coherence improvements, as well as their trade-off. To isolate the contributions of diffusion distance and directional coherence, we implemented two variants of CoDiCE: one utilizing classical weighted L_1 distance (CoDiCE$_{L_1}$) and another employing diffusion distance (CoDiCE$_{\text{diff}}$), both integrating directional coherence. The results are presented in Tables 1 3. In the paper, we employed ablation experiments as well, zeroing each term of formula (*) separately. Hyperparameter values for both algorithms were set based on preliminary grid searches and pilot experiments designed to balance sparsity, coherence, and feasibility. The final chosen hyperparameters $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$ reflected a practical trade-off between competing objectives, confirmed via ablation studies Publication I, Section 5.3.

Table 1: Evaluation metrics comparison for datasets with continuous features (**Domnich** et al. 2024).

Dataset	Metric	Validity ↑	Diffusion ↓	L1 continuous ↓	DCoherence ↑
Diabetes	CoDiCE$_{\text{diff}}$	**100%**	**0.38 ± 0.22**	1.11 ± 0.53	0.64 ± 0.15
	CoDiCE$_{L_1}$	**100%**	0.72 ± 0.45	**0.29 ± 0.16**	0.76 ± 0.16
	DiCE	54%	1.62 ± 0.73	1.10 ± 0.34	0.68 ± 0.13
	FACE	70%	1.64 ± 0.67	1.08 ± 0.36	0.72 ± 0.14
	Prototypes	26%	2.18 ± 0.88	2.12 ± 0.61	**0.84 ± 0.07**
	GS	**100%**	0.67 ± 0.31	0.39 ± 0.19	0.57 ± 0.12
Breast Cancer	CoDiCE$_{\text{diff}}$	60%	2.87 ± 1.21	2.18 ± 0.39	**0.78 ± 0.10**
	CoDiCE$_{L_1}$	60%	2.62 ± 1.07	1.22 ± 0.37	0.67 ± 0.09
	DiCE	46%	**2.13 ± 0.87**	**0.97 ± 0.28**	0.72 ± 0.08
	FACE	63%	2.91 ± 1.08	0.98 ± 0.38	0.74 ± 0.10
	Prototypes	31%	4.72 ± 0.45	2.22 ± 0.37	**0.79 ± 0.01**
	GS	**100%**	2.25 ± 1.31	0.47 ± 0.28	0.58 ± 0.08

The key empirical findings of this benchmarking are the following. CoDiCE consistently outperformed other methods in generating valid counterfactual explanations. For instance, on the Diabetes dataset, CoDiCE achieved 100% validity, significantly surpassing methods like DiCE (54%) and FACE (70%). Similarly, on challenging datasets such as German Credit and COMPAS, CoDiCE maintained validity (100%), highlighting its robust and effective search capability even under stricter feasibility constraints. For instance, even though prototype-based explanations had high coherence, they suffered low validity, 18% (few of their suggestions actually changed the prediction).

As anticipated, each algorithm naturally minimized the distance metric for which it was explicitly optimized. We observe that CoDiCE$_{\text{diff}}$ reports consistently lower diffusion distances, indicating that the generated explanations remained within well-connected regions of the data distribution. For continuous features in Table 1, for instance, the Diabetes dataset, CoDiCE$_{\text{diff}}$ significantly improved upon DiCE, FACE, Prototypes (Wilcoxon signed-rank test with all $p < 0.013$), while achieving comparable performance to GS ($p = 0.31$). For the L1 continuous distance metric,

Table 2: Evaluation metrics comparison across different frameworks for mixed types features datasets (**Domnich** et al. 2024).

Dataset	Metric	Validity ↑	Diffusion ↓	L_1 cont ↓	L_0 cat ↓	DCoherence ↑
Adult	$\text{CoDiCE}_{\text{diff}}$	**98%**	**0.001 ± 0.004**	3.9 ± 1.4	0.2 ± 0.1	0.84 ± 0.09
	CoDiCE_{L_1}	92%	0.005 ± 0.013	1.1 ± 0.6	0.4 ± 0.1	0.82 ± 0.08
	DiCE	78%	0.005 ± 0.011	1.2 ± 0.6	**0.1 ± 0.1**	**0.98 ± 0.02**
	FACE	82%	0.007 ± 0.013	**0.7 ± 0.3**	0.5 ± 0.2	0.81 ± 0.11
	Prototypes	19%	0.013 ± 0.012	1.6 ± 0.4	0.7 ± 0.1	0.85 ± 0.05
German	$\text{CoDiCE}_{\text{diff}}$	**100%**	**4.3 ± 3.4**	1.2 ± 0.5	**0.1 ± 0.1**	0.93 ± 0.04
	CoDiCE_{L_1}	**100%**	6.4 ± 3.8	**0.7 ± 0.4**	**0.1 ± 0.1**	**0.94±0.04**
	DiCE	49%	8.2 ± 3.2	1.2 ± 0.4	0.5 ± 0.1	0.78 ± 0.07
	FACE	63%	7.6 ± 2.4	1.0 ± 0.4	0.5 ± 0.1	0.79 ± 0.07
	Prototypes	34%	10.1 ± 3.4	1.1 ± 0.6	0.7 ± 0.1	0.73 ± 0.05
Compas	$\text{CoDiCE}_{\text{diff}}$	**100%**	0.03 ± 0.05	5.2 ± 1.9	**0**	0.92 ± 0.08
	CoDiCE_{L_1}	**100%**	0.03 ± 0.05	**0.8 ± 0.4**	**0**	**0.93 ± 0.07**
	DiCE	49%	0.03 ± 0.04	1.0 ± 0.6	0.4 ± 0.2	0.83 ± 0.11
	FACE	18%	0.04 ± 0.06	1.2 ± 0.6	0.5 ± 0.2	0.79 ± 0.11
	Prototypes	18%	**0.01 ± 0.01**	1.3 ± 0.7	0.6 ± 0.1	0.76 ± 0.09

Table 3: Evaluation metrics comparison across different frameworks for the Energy consumption dataset (regression task) (**Domnich** et al. 2024).

Dataset	Metric	Validity ↑	Diffusion ↓	L_1 continuous ↓	DCoherence ↑
Energy	$\text{CoDiCE}_{\text{diff}}$	100%	0.005 ± 0.03	0.52 ± 0.33	**0.67 ± 0.04**
	CoDiCE_{L_1}	100%	**0.003 ± 0.02**	**0.41 ± 0.26**	0.63 ± 0.11
	DiCE	100%	1.64 ± 2.21	0.51 ± 0.47	0.62 ± 0.11

CoDiCE_{L_1} significantly outperformed all baseline methods (Wilcoxon $p < 0.001$) except GS, where the difference was not statistically significant ($p = 0.089$). In contrast, directional coherence analyses showed that both CoDiCE variants significantly surpassed GS on the Diabetes dataset, indicating better alignment with human-intuitive feature changes (0.64 ± 0.15 vs. 0.57 ± 0.12, $p < 0.001$).

Across mixed-feature datasets (Adult Income, German Credit, COMPAS; Table 2), CoDiCE variants consistently demonstrated higher validity and improved proximity compared to established methods. However, diffusion distance can be lower when other algorithms find less valid counterfactual explanations, resulting in lower validity. For example, on the COMPAS dataset, the prototype-based method obtained a diffusion distance of 0.01 ± 0.01 (slightly lower than CoDiCE's 0.03 ± 0.05). Still, its validity was significantly lower (18% vs. 100%), suggesting that this seemingly better performance in distance was partly due to the failure to generate valid counterfactuals.

Regarding directional coherence, CoDiCE consistently showed higher DCoherence scores, suggesting that its proposed feature changes aligned closely with expected causal directions and model-based feature effects. For example, if a feature generally had a positive effect on the outcome, CoDiCE was unlikely to suggest decreasing it as part of the counterfactual. Methods that did not explicitly incorporate directional coherence (e.g., DiCE) occasionally suggested contradictory feature changes, reflected in their comparatively lower coherence scores. For example, on the Diabetes dataset, some DiCE-generated counterfactuals proposed to decrease insulin level while simultaneously increasing BMI, a combination counterintuitive to domain experts. Interestingly, prototype-based and FACE methods occasionally exhibited higher directional coherence scores compared to other baselines, likely due to their implicit constraints (such as using autoencoders or graph-based approaches) ensuring proximity to realistic data points. This observation suggests that directional coherence is indeed an independently important metric to assess the interpretability and practical usefulness of counterfactual explanations.

Finally, for the regression task predicting energy consumption, both CoDiCE variants demonstrated full validity (100%). They significantly improved upon DiCE in terms of diffusion distance (0.005 ± 0.03 and 0.003 ± 0.02 vs. 1.64 ± 2.21, Wilcoxon p-values: 0.025 and 0.001, respectively). CoDiCE_{L_1} showed significantly better diffusion distances than $\text{CoDiCE}_{\text{diff}}$ ($p = 0.016$). Regarding L1 continuous distance, both CoDiCE variants significantly outperformed DiCE (0.52 ± 0.33 and 0.41 ± 0.26 vs. 0.51 ± 0.47, $p < 0.04$). Additionally, both CoDiCE variants have higher directional coherence compared to DiCE (0.67 ± 0.04 and 0.63 ± 0.11 vs. 0.62 ± 0.11, $p < 1e-12$), with $\text{CoDiCE}_{\text{diff}}$ slightly outperforming CoDiCE_{L_1} ($p = 0.04$).

3.4. Summary and implications

By introducing diffusion distance as part of the counterfactual objective function, we redefine the notion of proximity, making it more compatible with realistic feasibility. This is particularly important in scenarios requiring multiple incremental steps for realistic, actionable changes. Many real-world changes are incremental or constrained. For example, improving one's health status involves gradual lifestyle changes, and altering creditworthiness takes a series of financial decisions over time. A counterfactual that suggests a dramatic one-shot change (jumping to a very different profile that a person cannot reach, such as participating in IRONMAN) is of little practical use. By modeling proximity through data connectivity, diffusion distance guides explanations along feasible, step-by-step paths observed in real data, which a person or system could actually take. This makes the resulting explanations actionable. Therefore, enhancing trust by demonstrating that the recommended changes are genuinely reachable, as they reflect the trajectories observed in real data.

Directional coherence is particularly crucial in domains where human intuition about feature influence is well defined (such as finance, healthcare, or legal decisions). In these settings, users have strong expectations about which factors should increase or decrease an outcome. Counterfactual explanations that run contrary to such expectations (for instance, recommending a decrease in income to get a loan approved) can lead to confusion or distrust in the AI system. By ensuring that the explanation logic is consistent with known cause-and-effect relationships, we make AI recommendations more acceptable and persuasive to humans. To achieve this, users can explicitly define constraints that represent partial causal knowledge, which take precedence in the counterfactual search. In cases where such constraints are absent, directional coherence defaults to using learned marginal feature influences. This cognitively inspired bias adds a layer of reliability, as stakeholders can follow the reasoning without feeling that the model is behaving oddly or against common sense. Additionally, directional coherence can act as an independent metric. The search for effective qualitative metrics is an important avenue of research, as the commonly used list of metrics is limited and does not reflect complex needs for good explanations (Keane et al. 2021b).

While embarking on the road to enhancing counterfactual explanations algorithms, we recognize several limitations. Calculating diffusion distance requires access to training data for building a representative diffusion space and may present computational challenges. However, in many real-world scenarios, particularly when explanations are generated by the same entity responsible for model training, having access to the training data is typically feasible and justifiable. Examples include internal audits, debugging model decisions by data scientists, or regulated environments (e.g., healthcare, finance) where full data transparency is legally required or operationally standard. As for computational complexity, diffusion distance is calculated once over the entire training set and does not significantly influ-

ence inference time. Second, we acknowledge that our choice of evaluation metrics is biased by terms in the objective function. Overall, the field currently lacks standardized evaluation practices, making algorithmic effectiveness commonly assessed through quantitative metrics tailored to optimization terms (Guidotti 2022; Kanamori et al. 2020; Verma et al. 2020). The best practice for evaluating counterfactual explanations is to conduct thorough user studies alongside quantitative evaluation metrics (Keane et al. 2021a). However, such user-centered evaluation might unfairly advantage interactive algorithms capable of adapting to user preferences, presenting a context-dependent bias. While the algorithm might prove useful for respondents, it still represents a biased subpopulation. Therefore, in our contribution, we focused on the algorithmic guarantee of effectiveness, leaving counterfactual evaluation to the next contribution.

The primary practical impact of our algorithm has been demonstrated through real-world deployments within the TRUST-AI (Transparent, Reliable and Unbiased Smart Tool) platform, developed as part of the Horizon EU TRUST-AI project. Feedback from industry partners in the energy and online retail sectors indicates that counterfactual explanations incorporating diffusion distance and directional coherence provided valuable and actionable insights for model developers and decision-makers. The CoDiCE framework is publicly available on GitHub and can be installed as a Python library: `https://github.com/anitera/CoDiCE`.

4. EVALUATING COUNTERFACTUAL EXPLANATIONS

4.1. Introduction

Although counterfactual explanations have become increasingly popular due to their contrastive and actionable nature, evaluating the quality of counterfactual explanations remains a significant challenge in the Explainable AI community, primarily due to the lack of standardized benchmarks and the complex, subjective nature of explanatory qualities. To date, hundreds of algorithms have been developed to generate counterfactuals, yet systematic evaluation practices are still lacking (Keane et al. 2021a). Quantitative metrics (validity, proximity, sparsity) overlook key human-centered aspects (Zemla et al. 2017), emphasizing algorithmic characteristics rather than user perspectives. While conducting user studies is considered the gold standard, their high cost and time-consuming nature create significant barriers to consistent and scalable evaluations. Moreover, different subpopulations, expertise levels, temporal differences, and visualizations lead to biases and inconsistent findings across studies, as well as reproducibility issues.

To address these challenges, this chapter introduces the CounterEval dataset, a publicly available benchmark capturing human judgments of diverse counterfactual explanations along multiple explanatory dimensions, including feasibility, trust, consistency, completeness, complexity, fairness, understandability, and overall satisfaction. Using this dataset, we explore two central Research Questions:

1. **RQ2a** Can overall satisfaction with counterfactual explanations be predicted from other explanatory qualities, and how do respondent backgrounds affect such assessments?
2. **RQ2b** Can we mimic human evaluative judgments of counterfactual explanations using Large Language Models (LLMs)?

4.2. Motivation

The challenge of evaluating explanation quality is common for the entire field of XAI. Many researchers emphasized the need for better evaluation frameworks (R. R. Hoffman et al. 2018, 2023; Longo et al. 2024; Rosenfeld 2021; Vilone et al. 2023). For example, Longo et al. (2024) stresses the lack of standardized assessment methods for explanations. Traditional quantitative metrics may have theoretical advantages, but often do not translate into what humans consider a "good" explanation. In practice, an explanation that is simply sparse or proximate to the original input is not necessarily useful or understandable to a user. The efforts of creating benchmarking datasets in XAI have mostly focused on feature attribution or rule-based explanations, with little emphasis on counterfactual explanations. For instance, Liu et al. 2021, XAI-BENCH is a collection of synthetic datasets

created to benchmark feature attribution methods, or a more recent extended version Sithakoul et al. 2024, Beexai. There are various natural language processing benchmarking datasets, such as commonsense question answering (Rajani et al. 2019, CoS-E), or causal explanation benchmarks (Abraham et al. 2022, CEBaB), or counterfactual explanations datasets (Kaushik et al. 2019, CAD), however, later is designed to evaluate model robustness or causal understanding through counterfactual training data rather than benchmarking explanation methods themselves.

To create a reliable benchmark, one should follow a gold standard of conducting evaluation via user studies. However another issue is even when user studies are conducted (only 7% of papers according to Keane et al. 2021b), they ask only cumulative metric of satisfaction or trust, which gives little insight what exactly led to high scores in one or the other explanation, missing essential explanatory virtues like consistency, feasibility, or fairness (Warren et al. 2023). This chapter aims to highlight that capturing these varied "explanatory virtues" is essential for user-aligned evaluation. Additionally, post-hoc explanation methods inherently intertwine with the performance of the underlying predictive models, complicating the assessment of the explanatory algorithm's efficacy independently from model accuracy. Consequently, human evaluations can merge intuitions about model correctness and explanation clarity, which requires carefully designed benchmark scenarios capable of isolating these aspects.

An additional motivation is the practical need to scale user evaluations, given their inherent cost and resource requirements. Here, we explore if the growing capabilities of LLMs might mitigate some of the constraints and replicate human judgments.

The need for better evaluation frameworks in XAI has been emphasized by many researchers (Keane et al. 2021a) (such as claiming that humans prefer simple, sparse explanations, conflicting with completeness and other findings where people prefer more complex explanations)

4.3. CounterEval dataset for benchmarking counterfactual explanation evaluation (Publication II, III)

To address the challenge of evaluating and comparing multiple counterfactual explanation algorithms and better understand the nature of human satisfaction with such explanations, we designed a diverse evaluation dataset of counterfactual scenarios **CounterEval dataset** (Domnich et al. 2024). First, we identified seven dimensions relevant for assessing counterfactual quality: *Feasibility*, *Consistency* (*Coherence*), *Completeness*, *Trust*, *Understandability*, *Fairness*, and *Complexity* (see Table 4). The selection of these dimensions is grounded in cognitive science literature on explanation quality (Keil 2006; Miller 2019; Zemla et al. 2017, and others) and explained in detail in Publication II, Section "Dimensions of Explanatory Qualities".

We carefully constructed 30 counterfactual scenarios explicitly varied across

Metric and scale	Description
Satisfaction from 1 to 6	This scenario effectively explains how to reach a different outcome
Feasibility from 1 to 6	The actions suggested by the explanation are practical, realistic to implement and actionable
Consistency from 1 to 6	All parts of the explanation are logically coherent and do not contradict each other
Completeness from 1 to 6	The explanation is sufficient in explaining the outcome
Trust from 1 to 6	I believe that the suggested changes would bring about the desired outcome
Understandability from 1 to 6	I feel like I understood the phrasing of the explanation well
Fairness from 1 to 6	The explanation is unbiased towards different user groups and does not operate on sensitive features
Complexity from -2 to 2	The explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex

Table 4: Definitions of the evaluation criteria provided to the respondents in the questionnaire with ranking scale (**Domnich** et al. 2025a).

these identified explanatory dimensions. To isolate the effects of individual metrics, most counterfactual explanations were generated using counterfactual explanation algorithms (DiCE and CoDiCE), applied to standard benchmarking datasets (Adult and Pima Indians Diabetes) commonly used for counterfactual evaluation. However, to produce more extreme and illustrative cases for specific dimensions (such as incompleteness or clear incoherence), some scenarios were manually crafted. This approach allows participants to concentrate on assessing the quality of explanations without being distracted by potential confusion over the features or underlying model details. To cover every explanation quality sufficiently, several sub-cases were considered, taking inspiration from the body of literature and a nuanced understanding of metrics (R. M. Byrne 2023; R. R. Hoffman et al. 2023):

- *Feasibility* was manipulated by varying the magnitude of feature changes, such as extreme or abnormal changes in continuous features (e.g., a large increase in income from 1,000 to 1,000,000) or categorical features (e.g., abrupt jumps from "school" to "PhD"). We also considered the difficulty of changing certain baseline values (e.g., changing from zero) to illustrate practical challenges in implementing certain counterfactuals. Explanations that do not meet this criterion have been shown to receive low ratings in empirical studies (Butz et al. 2024; McCloy et al. 2000).
- *Consistency (Coherence)* was ensured by constructing scenarios where feature changes either positively correlated or deliberately contradicted each other, following the notion of internal coherence as defined by Zemla et al.

(2017). For instance, scenarios were created where two positively related features (such as hours worked per week and income) either both increased (coherent) or moved in conflicting directions (incoherent). Both continuous and categorical features were systematically tested in various combinations.

- *Completeness* was managed by deliberately designing explanations to include varying numbers of features necessary for an outcome change, ranging from minimal to fully comprehensive causal chains. In some scenarios, intentionally incomplete explanations (omitting crucial intermediate steps) were included to assess user sensitivity to missing causal links and the necessity of contextual domain knowledge. Given that people often fill logical gaps based on context (Strickland et al. 2011), we preferred to omit multiple intermediate steps rather than just one, in order to assess this compensatory reasoning better.
- *Complexity (Simplicity/Sparsity)* was controlled by varying the number of features modified in each explanation (R. R. Hoffman et al. 2018; Ramon et al. 2021). We constructed explanations with different levels of complexity, with some involving very few changes (simple/sparse) and others involving multiple interconnected feature changes (complex).
- *Trust* was indirectly addressed by scenarios with varying degrees of plausibility, achieved by designing counterfactuals with clearly actionable and realistic outcomes, versus scenarios suggesting unrealistic or improbable outcomes. Following Stepin et al. (2022), we evaluate the perceived credibility of suggested changes.
- *Fairness* was varied by introducing sensitive or ethically controversial features (such as age, gender, or marital status) into certain explanations, allowing us to assess users' reactions to counterfactuals perceived as biased or discriminatory. While fairness has traditionally been approached as a quantitative metric Ge et al. 2022, there remains limited understanding of how it shapes users' perceptions of explanation quality.
- *Understandability* was not intentionally varied, serving instead as a comprehension check, given that not all participants were native English speakers. This dimension is also referred to as Readability Stepin et al. 2022 or Comprehensibility Ali et al. 2023; Vilone et al. 2021 in the literature. To maintain consistency, we deliberately avoided overly complex wording, ensuring that all scenarios were uniformly clear and accessible. Ensuring participant understanding of the explanation was essential for evaluating more intricate dimensions of explanation quality.

Another important consideration was that all counterfactual scenarios suggested positive directional changes from the original factual state, given that directionality influences users' perceptions (Kuhl et al. 2023). An example of a questionnaire question can be found in Supplementary materials of **Domnich** et al. (2025a): Appendix A, Table A.1.

Data collection proceeded in three phases following ethical committee approval. Initially, a pilot study with 15 respondents provided critical feedback regarding survey length and content clarity. Given that the pilot survey took approximately one hour to complete or sometimes even longer, we improved the questionnaire by removing excessive text and simplifying definitions and scenario descriptions based on the feedback. For instance, the metric initially named *Coherence* was renamed to *Consistency* to avoid ambiguity and mixing different interpretations.

The main data collection phase involved an initial group of 100 respondents, followed by a filtering step and a subsequent final phase involving 106 additional respondents, bringing the total to 206 respondents. Each respondent evaluated the scenarios according to the eight predefined evaluation metrics, including *Overall Satisfaction*. Each metric was measured on Likert scales from 1 to 6, except *Complexity*, which was rated from -2 (too simple) to 2 (too complex). The 6-point scale was chosen based on evidence that participants often do not meaningfully distinguish between adjacent points on larger scales (e.g., 7-point or more), leading to blurred responses and reduced data precision (Aybek et al. 2022). Moreover, an even number of response options (which forces a lean toward either side) was used to avoid the common tendency for participants to choose a neutral midpoint as a cognitive shortcut, a behavior described by Krosnick (1991) and well known in cognitive science, as it provides an easy way out to minimize effort. For data quality, participant eligibility criteria required respondents to be at least 18 years old and fluent in English. Demographic data includes age range, education level, and prior experience with machine learning, counterfactual explanations, or medical background.

The resulting dataset, named CounterEval, consists of evaluations across the eight metrics by 206 respondents and is publicly available on HuggingFace (`https://huggingface.co/datasets/anitera/CounterEval`) and Zenodo (Domnich et al. 2024), including demographic profiles of participants.

4.4. Human Rating Patterns of Counterfactual Explanations (Publication II, III)

A thorough data quality analysis was conducted to filter out random responses, a common issue in online surveys. Responses were reviewed based on a set of exclusion criteria: failing attention checks, unusually short response times (average response time ~42 minutes), an excessively low average *Understandability* score, significant clustering of ratings (suggesting repetitive or non-discriminative responses), and inconsistent answers to specific indicator questions (**Domnich** et al. 2025a Appendix A, Table A.2). Upon analyzing the responses, 10 participants who failed three or more of these criteria were excluded from the dataset. After filtering out incomplete responses, we obtained a final set of 196 valid respondents, each of whom evaluated all 30 scenarios, yielding a total of 5,880 ratings in our analysis (30 scenarios × 196 participants).

The analysis revealed that participants utilized the full range of scales, and scenarios elicited a broad spread of scores, indicating substantial variance in perception. The mean ratings per metric ranged from 3.0 to 4.8, with substantial standard deviations, indicating substantial variability in human judgments (further details are available in Publication II, Table 2). Figure 5 visually summarizes the distribution of responses for each explanatory dimension, categorized by *Overall Satisfaction* level divided into three classes (low, medium, high), further showcasing the spread of ratings. This visualization demonstrates the substantial variability and nuanced differences in human evaluations across diverse scenarios and explanatory criteria. Therefore, these results highlight the potential of the CounterEval dataset as a practical resource for benchmarking multiple aspects of counterfactual explanations.

Another significant observation was the strong inter-metric correlation. Figure 6 shows that all metrics positively correlate (Pearson p-value $< 10^{-4}$, after Bonferroni correction for 28 pairs), except for the pair *Complexity* and *Fairness*, which was the only non-significant association. The lack of correlation between complexity and fairness may suggest that scenarios involving fairness did not vary in complexity. Further analysis, as reported in Publication II, Figure 1, revealed that most explanatory metrics tend to rise or fall together across scenarios. This implies that raters who rated an explanation highly in one quality often rated it highly in others, treating explanation quality holistically.

Furthermore, we performed an exploratory factor analysis to uncover latent structures in the evaluation metrics. The Kaiser–Meyer–Olkin (KMO) measure was 0.893, confirming sampling adequacy for factor analysis. The Bartlett test was highly significant ($\chi^2 = 28{,}150.28$, $p < 0.001$), confirming that the correlation matrix is not an identity matrix. A three-factor solution emerged based on the eigenvalue "elbow" observed after the third factor, explaining approximately 65.9% of the variance in ratings. The first factor accounted for about 40.5% of the variance and showed high loadings from *Feasibility*, *Consistency*, and *Trust* (all loading coefficients above 0.75), suggesting a dominant dimension related to the practical plausibility and reliability of the explanation. The second factor explained an additional 16.4% of the variance and was mainly defined by *Understandability* (high loading of ∼0.78) and, to some extent, *Complexity*, which loaded negatively on this factor in the rotated solution. This indicates a dimension related to clarity or cognitive complexity. The third factor contributed a further 9.0% of the variance and showed moderate loadings from metrics such as *Completeness* and *Fairness* (loadings ∼0.57), suggesting a smaller yet distinct dimension possibly related to the breadth and ethical framing of the explanation. This factor structure reinforces that while many metrics co-vary (factor 1 grouping actionability and trustworthiness), certain aspects like complexity form an independent axis of quality.

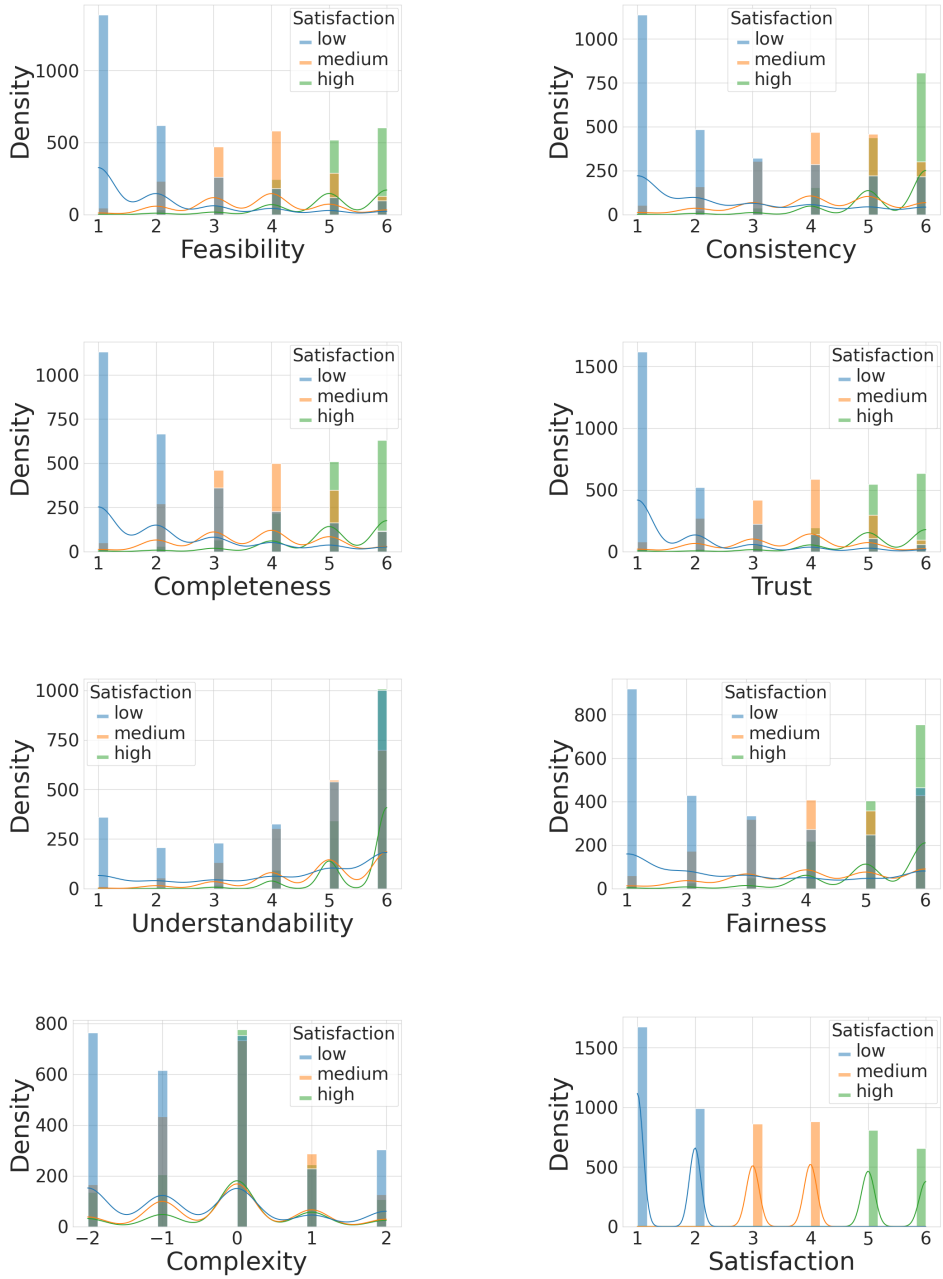


Figure 5: Per-metric distributions grouped by Satisfaction level (low, medium, high). Each histogram is color-coded by the participant's Satisfaction category, illustrating how individual metrics vary across these categories (**Domnich** et al. 2025b).

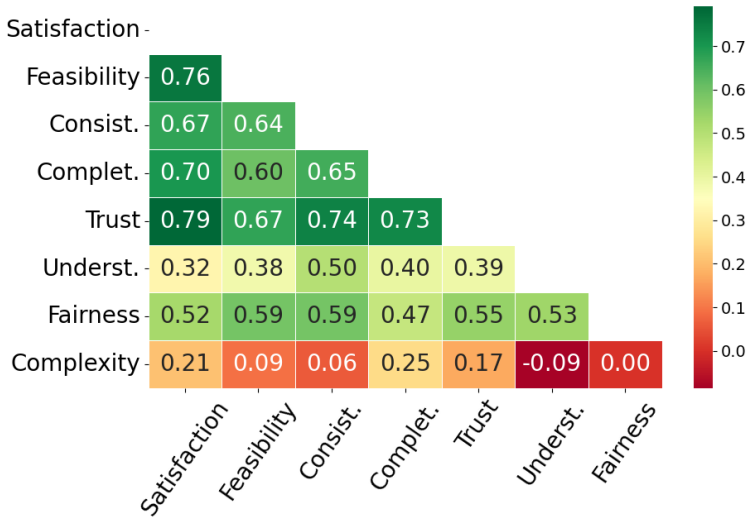


Figure 6: Spearman correlation table between metrics. The values for *Complexity* were mapped linearly from the original [-2,2] scale to [1,6] to be in line with the other metrics (**Domnich** et al. 2025a).

4.5. Predicting Overall Satisfaction from Explanatory Metrics (Publication II)

In Publication III, we continued data analysis, this time delving deeper into the relation of background data with participants' evaluation patterns and explored how well overall satisfaction with counterfactual explanations can be predicted from other explanatory qualities (*Feasibility*, *Trust*, *Consistency*, *Completeness*, *Complexity*, *Fairness*, and *Understandability*).

We conducted exploratory bi-clustering to uncover latent participant and scenario patterns. Clustering revealed distinct groups of respondents (highly critical vs. generally satisfied raters) and scenarios (consistently rated high or low across multiple dimensions). Factor analysis suggested a latent structure with *Feasibility*, *Trust*, *Completeness*, *Consistency*, and *Fairness* loading strongly onto one factor. *Complexity* emerged separately, reflecting independent perceptions regarding explanatory depth.

We found participants' backgrounds significantly influenced satisfaction ratings, with medical and machine learning expertise resulting in distinct evaluation patterns (p-values: medical experience, 0.0162; machine learning experience, 0.0299). Age and education showed no statistically significant effect across bi clusters ($p > 0.05$).

A central question is which of these explanatory qualities most strongly drives users' overall satisfaction and whether we can predict satisfaction from these dimensions. To answer this, we modeled Overall Satisfaction as a function of the seven other metric ratings. We applied a standard multiple regression analysis,

treating satisfaction scores (from 1 to 6) as a continuous outcome. The dataset cube of 30 counterfactual explanations, 196 evaluations, each providing ratings on seven explanatory metrics, resulting in 5880 total instances. For modeling, we considered two data-splitting strategies: random split and question split. Random split considers each question-metric pair independently and divides into a train/test split randomly (80%/20%). In scenario-based split (question split), we assign the entire scenario to either the training or test set, resulting in 24 scenarios (24×196 = 4,704 instances) for training and 6 scenarios (6×196 = 1,176 instances) for testing.

First, we applied Ordinary Least Squares (OLS) regression, which achieved a high explanatory power ($R^2 = 0.757$, 5-fold cross-validated ±0.008). This indicates that the selected explanatory metrics capture a substantial variability in overall satisfaction. In other words, the chosen explanatory metrics effectively capture most of what makes a counterfactual explanation satisfying to users. Table 5 reports coefficient estimates, where *Feasibility* (realism and actionability of the suggested changes) and *Trust* (belief in the effectiveness of the changes) emerged as the strongest predictors, each positively influencing satisfaction scores significantly ($\beta \approx 0.36$ each, $p < 0.001$). Completeness also significantly contributed ($\beta = 0.17$, $p < 0.001$). Although its coefficient is about half the magnitude of Feasibility or Trust, Completeness still contributes meaningfully: users do appreciate when an explanation provides a fuller picture, as long as it remains feasible and trustworthy. The other metrics have smaller but still statistically significant coefficients: Consistency ($\beta = 0.07$, $p < 0.001$) and Complexity ($\beta = 0.08$, $p < 0.001$) both have mild positive effects on satisfaction. These indicate that users prefer explanations that are logically consistent with the scenario and perhaps find slightly more complex explanations (richer in detail) more satisfying than overly simplistic ones, as long as complexity does not compromise clarity. The only metric that showed a negative coefficient in the multivariate model was Understandability ($\beta = -0.07$, $p < 0.001$). This result is expected as Understandability served as a comprehensibility check, and we expect a high Understandability score even for an incoherent explanation. Finally, Fairness had a very small positive coefficient ($\beta = 0.07$) that did not reach statistical significance ($p = 0.056$). This borderline result implies that, in our study, explicitly perceived fairness of an explanation had only a minor incremental effect on satisfaction once other factors were accounted for. Upon further analysis, we observed that participants did not fully agree on what constituted fairness in those contexts. Furthermore, we conducted a follow-up modeling excluding Feasibility and Trust. The model still explained 58% of variance, highlighting Completeness and Consistency as influential secondary factors. Notably, after excluding two criteria, all other metric coefficients became significant with *Completeness* ($\beta \approx 0.413$) becoming the strongest driver, followed by *Consistency* ($\beta \approx 0.322$), *Fairness* ($\beta \approx 0.182$), *Complexity* ($\beta \approx 0.096$), and *Understandability* remains negative ($\beta \approx -0.082$).

Table 5: OLS regression results modeling Overall Satisfaction. Reported are the coefficient estimates, standard errors (SE), t-values, p-values, and 95% confidence intervals (CI) for each predictor (**Domnich** et al. 2025b).

Predictor	**Coef.**	**SE**	**t-val**	**p-val**	**95% CI**
Intercept	0.1766	0.040	4.410	0.000	[0.098, 0.255]
Feasibility	**0.3581**	0.010	36.563	0.000	[0.339, 0.377]
Consistency	0.0665	0.010	6.347	0.000	[0.046, 0.087]
Completeness	0.1702	0.011	16.144	0.000	[0.150, 0.191]
Trust	**0.3618**	0.011	31.796	0.000	[0.340, 0.384]
Understandability	−0.0690	0.010	−7.036	0.000	[−0.088, −0.050]
Fairness	0.0170	0.009	1.908	0.056	[−0.000, 0.034]
Complexity	0.0802	0.010	7.658	0.000	[0.060, 0.101]

Comparative analysis of machine learning models (OLS Regression, Decision Tree, Random Forest) showed similar results, with Decision Trees performing slightly better on random splits but demonstrating lower generalization to unseen scenarios, suggesting potential overfitting. Converting satisfaction scores into three classes (low, medium, high), we applied Logistic Regression, Decision Tree, and Random Forest classification models. Decision Trees again exhibited the highest accuracy (78%) under random splits but lower stability on scenario-based splits (70%). Logistic Regression and Random Forest provided consistent, robust performance (75% accuracy) across both splits. Further analysis using SHAP values for less interpretable models, like Random Forest, indicated *Feasibility* and *Trust* as dominant factors influencing class predictions, although *Completeness*, *Consistency*, and *Fairness* were decisive for distinguishing between low and medium satisfaction levels.

4.6. Mimicking Human Judgment Using LLMs (Publication III)

After filtering participant responses based on criteria described in Section 4.4, we obtained averaged human ratings per scenario-metric pair, giving a data cube of 196 valid respondents for 30 counterfactual explanations with eight explanatory metrics. In the survey, each metric was initially rated on a Likert scale from 1 to 6, except *Complexity*. To simplify ratings and improve LLM generalization, these scores were grouped into three distinct categories: “low” (1–2), “medium” (3–4), and “high” (5–6). Given recent advancements in the capabilities of LLMs, we hypothesize that these models could reliably predict explanatory metric assessments provided by humans. Thus, we framed this as a supervised three-class classification task, in which an LLM, given a counterfactual explanation and the corresponding metric question, predicts the appropriate categorical rating.

Figure 7 summarizes our methodological approach. The dataset annotated by human respondents was partitioned into two types of experimental splits: metric-

based and scenario-based (question-based). In the metric-based split, each metric was equally represented in the test set, ensuring that LLMs could generalize across metrics. On the other hand, the scenario-based split reserved entire counterfactual explanations with their 8 metrics for testing, which checks the LLMs' ability to generalize to completely unseen explanations.

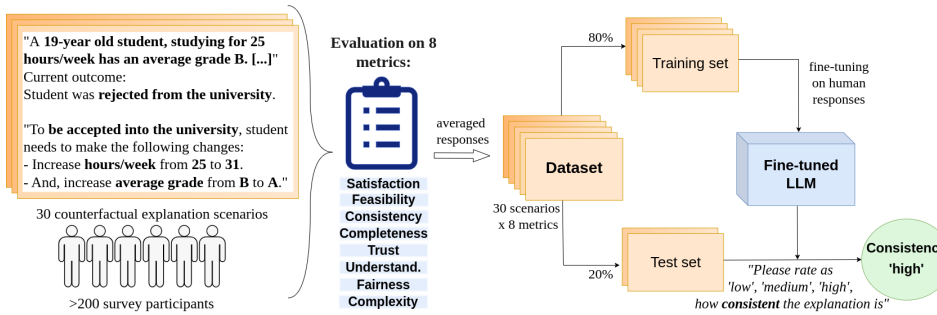


Figure 7: Pipeline for assessing LLMs' ability to mimic human evaluations using collected CounterEval dataset. Human respondents evaluated counterfactual explanations that varied across multiple explanatory dimensions. These evaluations were used to fine-tune and test several LLM models, comparing their predictions to human judgments on a reserved test set (**Domnich** et al. 2025a).

For the LLM evaluation and fine-tuning, we selected three models: Llama 3.1 Instruct, Llama 3 Instruct, and GPT-4. GPT-4 was accessed via the OpenAI API, while the Llama models were fine-tuned using the HuggingFace transformers library on HPC clusters equipped with NVIDIA Tesla A100 GPUs. To reduce memory usage during the fine-tuning process, we employed QLoRA. This method integrates quantization with low-rank adapters through rank decomposition matrices, which makes fine-tuning more efficient and feasible on available computational resources.

The prompts for LLM training and testing were designed based on the original human evaluation survey, consisting of two parts: system prompts and user prompts (instructions were directly adapted from a questionnaire, a pair of factual-counterfactual explanations). System prompts set the context for the LLM evaluation task. We tested several variations for the system prompt (baseline, enhanced). At the end, we used the prompt with an introduction to counterfactual explanations, the specific metric definition, and the expected output format. As for the user prompt, we directly adapted it from the questionnaire. It consists of a pair of factual-counterfactual scenarios, and a rating instruction tailored for each metric, e.g., "Please rate as 'low' (very infeasible), 'medium', or 'high' (completely feasible), how feasible is this explanation?"

Tables 6 and 7 present the averaged performance over four runs for predicting the averaged respondents' estimation of each explanatory quality. In a zero-shot scenario without fine-tuning, GPT-4 achieved an accuracy of 63%, significantly outperforming random guessing (33%), as confirmed by a binomial test

($p < 0.001$). However, fine-tuning improved model performance, with models such as Llama 3.1 (8B parameters) and Llama 3 (70B parameters) reaching accuracies up to 85%. Table 7 presents the performance of each metric individually for the best-performing Llama 3 70B model.

Model	**Metric Split**		**Question Split**	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Llama 3 8B	0.48	0.80	0.45	0.77
Llama 3.1 8B	0.52	**0.85**	0.50	0.74
GPT-4	**0.63**	-	0.58	-
Llama 3 70B	0.57	**0.85**	**0.59**	**0.81**

Table 6: Accuracy for metric-based and question-based testing set across evaluated LLMs. Scores averaged over 4 runs, highest score for each column highlighted in bold (**Domnich** et al. 2025a).

Furthermore, we tested the ability of fine-tuned LLMs to mimic the judgments of individual respondents. We clustered all responses in four clusters and selected a random representative from each cluster. Table 8 demonstrates that fine-tuned LLMs effectively captured individual rating patterns, achieving accuracies of up to 90%. This finding indicates the potential feasibility of developing personalized LLM-based evaluators. However, we observed variability across participants. Specifically, predictions for Participant B were less accurate, reaching only 66%. It is worth highlighting that baseline zero-shot predictions were also lower for this participant compared to the other three, suggesting inherent complexity or inconsistencies in this individual's rating patterns. This question brought an investigation of participants' clusters and a comparison of demographic data, which we discuss in the next section.

Metric	**Metric Split**		**Question Split**	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Satisfaction	0.67	**0.96**	0.50	**0.88**
Consistency	0.58	0.83	0.83	0.88
Feasibility	0.79	0.96	0.54	0.67
Understandability	0.54	**1.0**	0.92	1.0
Fairness	0.50	**0.83**	0.67	**1.0**
Trust	0.50	0.67	0.50	0.50
Complexity	0.42	**0.75**	0.42	**0.83**
Completeness	0.33	**0.83**	0.33	**0.75**

Table 7: Evaluation of various metrics for Llama 3 70B Instruct model. The largest improvements are highlighted in bold. Each of the accuracy scores is the average score over 4 runs (**Domnich** et al. 2025a)

Participant	Zero-shot	Fine-tuned
A	0.67	0.87
B	0.58	0.66
C	0.69	0.90
D	0.69	0.90

Table 8: Evaluation accuracy over all metrics for four participants that were selected to represent different subgroups of participants (**Domnich** et al. 2025a).

4.7. Summary and implications

This chapter contributes to the evaluation of counterfactual explanations by introducing **CounterEval**, a publicly available dataset consisting of over 200 human evaluations across multiple explanatory dimensions. It enables benchmarking and comparison of counterfactual algorithms to improve standardized assessments within the XAI community. The dataset is openly accessible for researchers and practitioners via HuggingFace (`https://huggingface.co/datasets/anitera/CounterEval`). We invite researchers to expand it further.

Using the CounterEval dataset, two central research questions were addressed. First, we demonstrated how *Overall Satisfaction* with explanations could be predicted from specific explanatory qualities such as *Feasibility*, *Trust*, *Completeness*, and *Consistency*, and how these evaluations differ across respondent backgrounds (Publication II). Second, we explored whether Large Language Models (LLMs) could reliably mimic human evaluative judgments (Publication III). Our findings emphasize the utility of these explanatory metrics in capturing nuanced human preferences, but also highlight that some dimensions, such as *Fairness* and *Complexity*, may require deeper conceptual clarification and operationalization in future studies.

A critical insight from this work is the necessity of incorporating multiple explanatory dimensions rather than relying solely on aggregate measures like *Overall Satisfaction*. Capturing diverse aspects such as *Completeness*, *Consistency*, *Complexity*, and *Feasibility* allows for a more nuanced understanding of what users value in explanations and can help guide the development of future counterfactual methods. However, our results also caution against oversimplified interpretations, especially given the variability observed across demographic and experiential backgrounds. The findings regarding demographic influences underscore the critical need to tailor explanatory methods and evaluations to specific user groups, rather than assuming universal evaluative standards.

Furthermore, the insights obtained regarding the relationships among explanatory metrics and their connection to overall satisfaction suggest avenues for developing improved quantitative evaluation methods. A better understanding of the importance of different explanatory qualities can inform future efforts in crafting more user-aligned evaluation metrics and enhancing the design of counterfactual explanation techniques. Finally, the demographic insights emphasize the necessity

of adapting explanatory methods to specific audiences.

Regarding RQ2b, our investigation demonstrates the potential of LLMs to complement human evaluation by mimicking average human judgments. Achieved performance with fine-tuned LLMs verify that they indeed can serve as scalable proxies for human assessment. However, we acknowledge that LLM-based evaluations should complement rather than replace human judgment entirely. The observed variability in predicting individual participant ratings emphasizes that LLM-based evaluations should be applied cautiously and critically, primarily in contexts where large-scale human evaluations are infeasible. Fine-tuned LLMs, publicly available on GitHub (`https://github.com/anitera/CounterEval`), can be used out of the box for evaluating a counterfactual algorithm if the output is transformed to a prompt.

Overall, these contributions underscore the need for comprehensive benchmarking datasets and robust evaluation frameworks capable of assessing counterfactual explanation methods independently from the predictive performance of the underlying models. It is worth noting that despite a relatively substantial sample size (200 respondents), our participant demographics may not fully represent diverse expert or cultural perspectives. For instance, domain experts in specialized fields or users from diverse cultural backgrounds might rate explanations differently (as suggested by our cluster analysis). Moreover, explicitly including diverse explanatory virtues in user studies provides valuable quantitative dimensions that can enhance future explanation evaluations.

5. COUNTERFACTUAL INPAINTING FOR MEDICAL IMAGING (PUBLICATION IV)

5.1. Introduction

This chapter addresses the third research question of this thesis, **RQ3**: *Can we apply counterfactual explanations to produce Weakly Supervised Semantic Segmentation labels effectively?*. Specifically, we explore the potential of counterfactual explanations in medical imaging for Weakly Supervised Semantic Segmentation (WSSS). We introduce a method called Counterfactual Inpainting (COIN), which uses a generative adversarial network to inpaint regions of the classifier's interest to flip the prediction from positive (i.e., tumour is present) to negative (i.e., removing the tumour). The method can assist in medical diagnostics, where providing interpretation for the model is crucial for trustworthiness. But also to assist segmentation by treating the difference between the original and counterfactual images as a weak supervision signal. The method is evaluated on both synthetic and Tartu University Hospital (TUH) real CT datasets and compared with several attribution methods and the primary counterfactual method.

5.2. Motivation

Motivated by the exploration of whether counterfactual explanations can be effectively extended beyond tabular data to address explainability challenges in complex, high-dimensional medical imaging data to support high-stakes model decisions, this investigation builds on several hypotheses. The first hypothesis is that counterfactual explanations can provide better explainability compared to commonly used methods like Class Activation Maps (CAMs), such as GradCam (Selvaraju et al. 2017), ScoreCAM (H. Wang et al. 2020), LayerCAM, RISE (Petsiuk et al. 2018), etc. Despite their widespread adoption, CAM-based approaches have significant limitations, as they typically highlight only the most discriminative image regions rather than capturing the entire object or pathology of interest (Jeanneret et al. 2022). Additionally, research by Ghassemi et al. (2021) emphasized that CAM methods can produce saliency maps that appear unchanged even when input changes significantly and alter model predictions, potentially misleading clinicians and offering a "false hope" of explainability rather than genuinely interpretable insights. Such failure cases can mislead clinicians and are unlikely to achieve the desired goals of trust and understanding of trained models. The contrastive nature of counterfactual explanations directly addresses the decision boundary: "if the feature were different, the diagnosis would flip". It helps to generate explanations with a clear contextual boundary. Another study pointed out that generating realistic counterfactual images led to better mental models, higher satisfaction, and greater trust among medical practitioners, compared to highlight-based explanations Mertes et al. (2022). Singla et al. (2023) pointed out that coun-

terfactual explanations tend to surface features that are causally relevant to the model's prediction, as they perform a virtual intervention on the image to test the model's reliance on a feature. This kind of insight is invaluable in medicine: it helps confirm whether the AI is focusing on true pathological signals (heart size, costophrenic angle) versus spurious correlations. Moreover, counterfactual explanations have an element of actionability or recourse, suggesting how a patient's data could be different for a favorable outcome. While not all suggestions are directly actionable (patients cannot simply change their anatomy), this format at least frames explanations in terms of concrete changes.

The second hypothesis is that if counterfactual explanations indeed provide reliable and meaningful interpretations, they could also serve effectively as weak labels for semantic segmentation tasks. WSSS is especially valuable in the medical imaging context, where dense, pixel-level annotations from domain experts are costly, time-consuming, and challenging to obtain. Medical scans, such as CT images, require meticulous pixel-level labeling, often involving multiple annotators to reach consensus due to the high-stakes nature of medical decisions. Given this resource-intensive process, the potential for automatically generating weak segmentation labels from high-quality explanations offers a significant practical advantage. Instead of beginning annotation from scratch, annotators could refine initial labels derived algorithmically, saving time and ensuring high-quality annotations. This hypothesis was supported by the initial findings of Singla et al. (2023), who demonstrated the feasibility of producing counterfactual explanations for the lung cancer dataset using GANs. Their work showed that generative models could create realistic alternative versions of medical images to explain classifier decisions. However, their approach required pre-existing segmentation masks during training, which is conflicting with applicability in weakly supervised settings where such masks are not available. Therefore, testing the possibility to generate counterfactual explanations that preserve realistic abnormal images, preserving all patient details, became the main challenge.

5.3. Counterfactual Inpainting Algorithm

The counterfactual inpainting approach is designed to produce weak segmentation masks from a pre-trained classifier using a generative adversarial network. Figure 8 illustrates the method. Given an abnormal medical image, the COIN pipeline first uses the classifier's prediction to identify regions indicative of pathology. The GAN then generates a counterfactual image by inpainting these pathological areas, altering the classifier's prediction from positive (pathological) to negative (normal). Computing the absolute difference between the original and the counterfactual image provides a weak segmentation label of the pathological regions without requiring explicit segmentation masks during training.

The method was developed starting from Singla et al. (2023) counterfactual approach, but employing several improvements. The main difference is that to

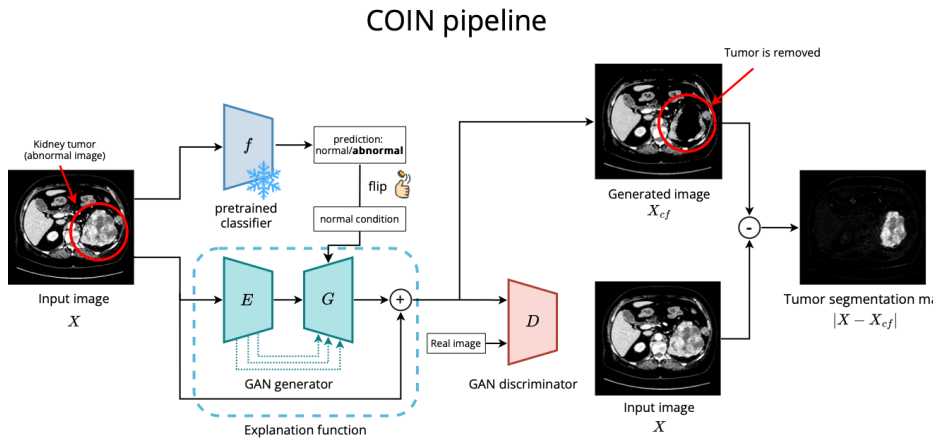


Figure 8: Overview of the counterfactual inpainting (COIN) pipeline. Given the input image X and black-box classifier f that produces a classification label, the image-to-image model (GAN) generates a counterfactual image X_{cf} with $y = 0$. If X is abnormal, it is expected that X_{cf} no longer contains the abnormal part of the input image. Computing the absolute difference of the original image X and counterfactual image X_{cf} results in a weak tumor segmentation map (Shvetsov et al. 2024).

apply counterfactual explanations for WSSS use, we have to eliminate the use of segmentation labels within the algorithm itself. Over a series of experiments, multiple changes were added to the method. COIN uses perturbation-based GAN integrated with skip-connections, which improved the realism of generated images. Skip-connections helped to preserve detailed spatial information by combining feature maps from earlier layers with later layers.

The training objective of the COIN framework combines multiple loss terms to ensure realism, classifier consistency, and spatial coherence in generated images. The complete objective function for training GAN in COIN can be expressed as:

$$L_{total} = \lambda_{GAN} L_{GAN} + \lambda_f L_f + \lambda_{idt} L_{idt} + \lambda_{tv} L_{tv}, \tag{5.1}$$

where each term addresses specific aspects of the counterfactual generation process.

The GAN loss term L_{GAN} ensures that the generated counterfactual images remain realistic and indistinguishable from real medical images. The classifier consistency loss L_f guarantees that generated counterfactual images effectively flip the classifier's predictions from abnormal to normal. The identity loss term L_{idt} preserves important structural details and ensures that minimal changes are introduced to the non-pathological regions. Compared to Singla et al. (2023), we removed the usage of segmentation labels for the identity term. Finally, the Total-Variation (TV) loss L_{tv} is added to promote spatial coherence and reduce unrealistic artifacts, resulting in smoother and medically plausible counterfactual regions.

The total-variation loss term was not used in earlier approaches and was one of our additions that considerably improved smoothness. The total-variation loss term is defined as:

$$L_{TV}(X_{cf}) = \sum_{i,j} |X_{cf}^{i+1,j} - X_{cf}^{i,j}| + |X_{cf}^{i,j+1} - X_{cf}^{i,j}| \tag{5.2}$$

5.4. Evaluation of Counterfactual Inpainting

We evaluated COIN using two datasets: the TotalSegmentator dataset, comprising diverse CT scans with synthetic anomalies inserted into kidney regions; and the TUH clinical dataset from Tartu University Hospital, consisting of 291 annotated kidney tumor cases and 300 control scans. Both datasets were split into training and validation subsets (80%/20%), with TUH dataset splits stratified by tumor size. Figure 9 provides qualitative comparisons of the segmentation results produced by different methods. COIN consistently outperformed traditional attribution methods such as ScoreCAM, LayerCAM, and RISE by accurately localizing pathological regions while significantly reducing false positives and missed detections in both synthetic anomaly images and real tumor scans. By contrast, CAM-based maps tend to miss parts of the abnormality (focusing only on the most salient portion), and even the enhanced Singla et al. (2023) method with extra postprocessing still produced lower fidelity outputs. COIN's inpainting gives clear difference maps that closely match the true abnormal regions, resulting in more accurate and cleaner segmentation masks.

Table 9: Metric results for the attribution methods and the proposed counterfactual inpainting pipeline on the TUH dataset. Since CAMs and RISE do not create counterfactual images, FID and CV metrics cannot be computed for these methods (Shvetsov et al. 2024)

Datasets	Methods	FID ↓	CV ↑	IoU ↑
TotalSegmentator	ScoreCAM	-	-	0.030
	LayerCAM	-	-	0.026
	RISE	-	-	0.397
	Singla et al.*	0.047	0.998	0.445
	COIN	**0.003**	**0.997**	**0.646**
Tartu University Hospital	ScoreCAM	-	-	0.293
	LayerCAM	-	-	0.296
	RISE	-	-	0.294
	Singla et al.*	0.203	0.992	0.352
	COIN	**0.036**	**0.980**	**0.432**

Table 9 compares COIN with traditional attribution methods (ScoreCAM, LayerCAM, RISE) and a re-implemented version of Singla's counterfactual approach.

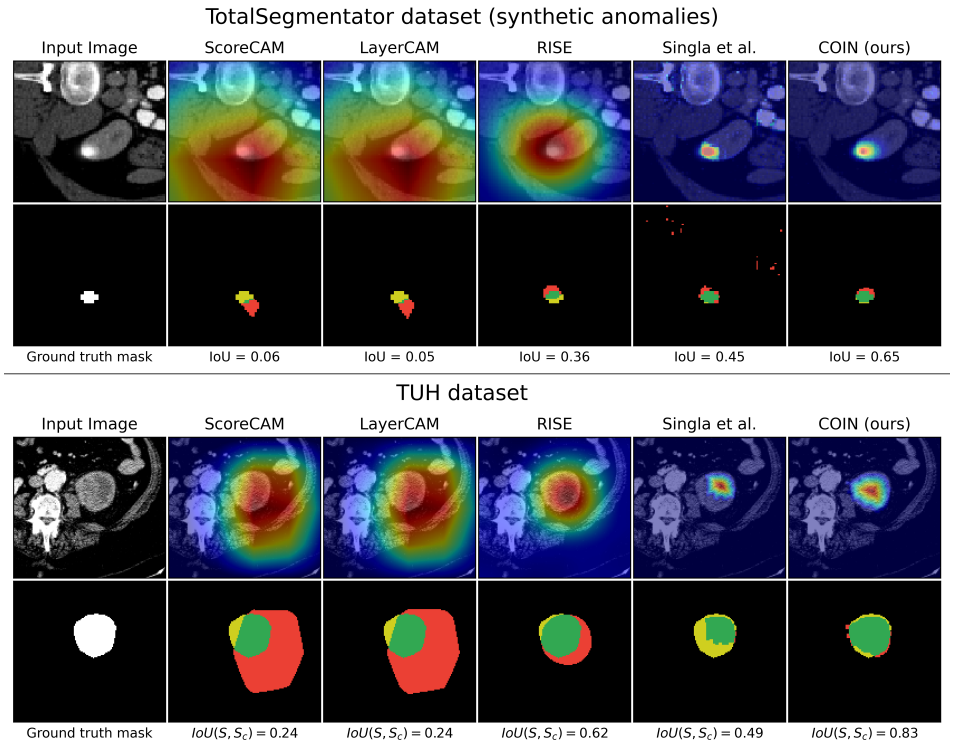


Figure 9: Visualization of the attribution and the counterfactual inpainting methods' predictions on TotalSegmentator and TUH datasets. For each dataset, the bottom row shows thresholded masks from saliency maps. For each mask, colors represent outcomes in terms of true positive (green), false positive (red), and false negative (yellow) predictions. Images are zoomed in for better clarity (Shvetsov et al. 2024).

On the TotalSegmentator synthetic dataset, COIN achieved an Intersection over Union (IoU) of 0.646, significantly outperforming the Singla et al. baseline (0.445) and RISE (0.397). COIN also produced the most realistic counterfactuals, with a Fréchet Inception Distance (FID) of 0.003 (where lower values are more realistic). On the real tumor dataset (CT scans from Tartu University Hospital), COIN achieved an IoU of 0.432, substantially higher than the attribution-based methods (which achieved IoUs around 0.29–0.30). The next best method (the enhanced Singla et al. counterfactual) reached 0.352 IoU. Counterfactual Validity (CV) scores for COIN were close to 1 (0.997 on synthetic anomalies and 0.980 on real tumors). This indicates that the generated counterfactuals almost always succeeded in flipping the classifier's prediction from "abnormal" to "normal", confirming that the inpainted regions effectively removed the pathological cues.

5.5. Summary and implications

This work presents a generative counterfactual explanation method for weakly supervised semantic segmentation in medical imaging called counterfactual inpainting. The COIN method builds on prior counterfactual algorithms by removing the need for segmentation masks during training and introducing architectural enhancements such as a perturbation-based generator, skip-connections in the GAN, and an added total-variation loss function term. The same counterfactual that explains "why the image was classified as abnormal" acts as a proxy segmentation mask of where the abnormality is. COIN demonstrated performance gains over existing attribution methods and a prior counterfactual approach that has access to segmentation masks themselves, indicating that counterfactual explanations can indeed serve a dual role as a form of weak supervision.

One important limitation to note is that the effectiveness of COIN depends on the classifier's performance, meaning that if the underlying classifier is poorly trained or biased, the counterfactual generator will produce irrelevant regions. However, such behavior can serve another purpose: to debug the classifier. During initial replication of the Singla approach for the lung pneumonia detection dataset, we observed the classifier with near-perfect performance. However, counterfactual explanations produced X-rays that highlight arrows and marker letters "R" in the corner of images. It turned out that the classifier, when trained on partial data, was biased toward different markers that are present in the images. Jeanneret et al. (2022) stresses that an effective visual explanation should consist of "semantically meaningful perturbations" instead of arbitrary noise. The causal nature of these explanations further increases user trust, as the reasoning becomes inspectable and verifiable. A doctor can look at the counterfactual result and agree that, yes, if the opacity in the X-ray were gone, the algorithm would have no reason to label it pneumonia. This alignment between the logic of the model and the expectations of the clinician builds confidence in AI. COIN was already used in a benchmark of WSSS methods for Fibrotic Lung Disease (Yue et al. 2024), which could lead to a broader adoption of counterfactual approaches by other fields.

Therefore, we highlight the relevance of counterfactual explanation for multiple tasks, and we plan to explore the generalizability of COIN to other domains, where obtaining pixel-level labels is challenging. This chapter demonstrates that counterfactual explanations are not limited to post-hoc explainability and they can actively contribute to model training, providing a weak supervision signal.

6. DISCUSSION

This thesis has explored the generation, evaluation, and application of counterfactual explanations through four distinct but interconnected studies. The main objective of the work is to align counterfactual explanations more closely with human preferences by incorporating explanatory virtues, such as *feasibility*, *coherence*, *completeness*, and others. In this discussion, we reflect on the broader implications of our findings, identify current limitations, highlight emerging trends, and opportunities for future research.

RQ1*: Can we extract more feasible and coherent explanations compared to existing methods?*. The ***first publication*** introduces the CoDiCE framework (described in Chapter III), which proposes a novel counterfactual search approach by integrating diffusion distance and directional coherence. Feasibility, as modeled through diffusion distance, captures the realism of counterfactual transitions by identifying paths that follow the data manifold. Rather than suggesting unrealistic one-shot changes, this approach proposes feasible step-by-step modifications that users might realistically act on. Directional coherence enforces alignment between the suggested feature changes and the feature trends learned by the model. By preserving these marginal trends, explanations stay closer to human causal expectations. In cases where the user provides partial causal knowledge, these user-defined constraints can override the marginal trends learned by the model. The objective function then prioritizes user constraints where available and defaults to marginal trends for features without explicit constraints. This is particularly important in domains like finance or healthcare, where explanations that conflict with intuitive causal reasoning can undermine trust. We evaluated CoDiCE by benchmarking it against several established counterfactual explanation algorithms, including DiCE, FACE, Guided Prototypes, and Growing Spheres, across multiple datasets with classification and regression tasks.

In the broader landscape of counterfactual explanation research, CoDiCE occupies a unique intersection between geometry-based feasibility and cognitively grounded coherence. Existing feasibility-focused approaches (e.g., FACE (Poyiadzi et al. 2020)) model the data manifold but remain agnostic to intuitive causal directionality, while coherence-aware approaches often rely on post-hoc plausibility checks. CoDiCE integrates both constraints directly into the optimisation process, aligning with emerging calls in XAI for algorithms that not only produce technically valid counterfactuals but also adhere to the explanatory norms humans expect (Longo et al. 2024). Notably, evaluation across classification and regression contexts is relatively rare in counterfactual explanations.

However, including multiple objectives simultaneously introduces potential trade-offs, as highlighted by ablation experiments. Given that there are multiple usages for generating explanations, such as model debugging or suggesting actions based on an explanation (as discussed in cognitive psychology and explanation theory (Keil 2006)), it is advisable to weight biases differently across explanatory virtues,

depending on the intended use. Future improvements could consider other dimensions of human preferences, including completeness, fairness, and trust, as well as incorporating uncertainty estimation and diversity. Multi-objective optimization strategies can be employed to simultaneously optimize multiple criteria and find Pareto-optimal solutions (Rasouli et al. 2024).

Directional coherence also proved to be an insightful metric for comparing methods, with approaches that maintain alignment with the data distribution consistently scoring higher. The search for evaluation metrics like directional coherence reveals broader gaps in the evaluation of explainability. Commonly used measures, such as validity, proximity, and sparsity, fail to capture the user-centered perception of satisfaction and trust. The field lacks standardized metrics to evaluate whether an explanation is good from a human-centric perspective. This recognition informed the motivation for creating the CounterEval benchmarking dataset and tackling the challenge of evaluation from different angles.

In response to these gaps, Chapter IV introduces a CounterEval dataset of over 200 human-annotated counterfactual assessments rated across eight explanatory dimensions. Compared to other surveys, in addition to collecting a perceived measure of overall satisfaction, we evaluated each counterfactual scenario with more explanatory metrics (feasibility, consistency, completeness, fairness, trust, understandability, complexity). This approach allows a deeper understanding of how different user groups approach counterfactual evaluation, supporting the view that explanation evaluation must be multidimensional (Nauta et al. 2023).

RQ2a*: Can overall satisfaction with counterfactual explanations be predicted using human ratings of explanatory metrics?*. In the ***second publication*** discussed in Chapter IV, we modeled overall satisfaction from other metrics and identified feasibility (the actionability of suggested changes) and trust (the belief that the changes would lead to the desired outcome) as the most important predictors of overall satisfaction. However, when feasibility and trust are fixed, other metrics play a significant role in explaining 58% of the variance. We also identified that certain demographic profiles, such as experience with machine learning or a medical background, significantly influence ranking patterns. These findings align with prior evidence that clinicians value more complex explanations, as they confirm their causal beliefs and consequently reinforce trust in the system (Barbu et al. 2025; Ghassemi et al. 2021). On the other hand, machine learning experts, who understand the internal workings of algorithms, tend to display a lack of trust in automation (Ehsan et al. 2024). This highlights a tension between designing general-purpose explanations and tailoring them to specific audiences. The dataset was deliberately constructed for general-purpose benchmarking and did not require domain-specific knowledge. This design ensures accessibility but limits direct generalization to specialized contexts, such as high-stakes medical or legal decision-making. Furthermore, participants were recruited as members of a general audience, and therefore did not reflect domain-specific expertise or culturally diverse perspectives. Moreover, it is important to acknowledge that al-

though the 30 scenarios evaluated covered diverse explanatory virtues, they may not fully capture the complete spectrum of real-world counterfactuals, particularly in specialized domains. Future research should replicate and extend the dataset for targeted domains, integrating culturally and professionally diverse participant groups to capture the heterogeneity of explanatory preferences more effectively.

Methodologically, our approach lays the groundwork for developing quantitative evaluation metrics that can predict overall satisfaction from more granular explanatory dimensions. Such derived metrics could enable more standardized and scalable evaluation of explanations while preserving sensitivity to human-centered qualities. Establishing quantitative measures remains an important avenue for future research, and the present dataset provides a foundation for these efforts.

RQ2b*: Can large language models (LLMs) reliably estimate the quality of counterfactual explanations by mimicking human judgment?.* Recognizing the difficulty of conducting user evaluation at scale, in the ***third publication*** discussed in Chapter IV, we attempted to evaluate these explanations using LLMs and fine-tune them to mimic average human judgment as well as individual nuanced perceptions, achieving 81%-90% accuracy depending on the split and setting. This approach opens a possibility to of complementing counterfactual explanation evaluation with LLM acting as an expert approach.

However, the approach raises important ethical and methodological considerations. LLMs may inherit or amplify biases present in human judgments, particularly when trained on unbalanced or non-representative datasets. For instance, our analysis found that human evaluators often correlate multiple explanatory metrics together (e.g., associating high feasibility with high trust) and judge explanation holistically, while non-fine-tuned LLMs tend to decouple them. Additionally, using LLMs without continuous monitoring or updating of training data can lead to explanations optimized merely to satisfy model preferences rather than genuinely improving relevance to human users.

Looking forward, as LLM capabilities continue to evolve, iterative retraining with diverse data can enhance their ability to accurately mimic human evaluations, and targeted fine-tuning for specific expert groups may further enhance their utility. Nevertheless, it remains essential to emphasize that LLM-based evaluation should complement, not replace, the nuanced insights provided by human evaluators, especially when direct human evaluation is not readily accessible.

RQ3*: Can we apply counterfactual explanations to effectively produce Weakly Supervised Semantic Segmentation labels?.* The ***fourth publication*** introduced in Chapter V pushes counterfactual explanations beyond tabular data and tests its application in the field of Weakly Supervised Semantic Segmentation. The COIN framework uses GAN-based counterfactual inpainting for generating segmentation masks from a pre-trained classifier. By flipping classification outcomes (e.g., from abnormal to normal) and measuring the visual delta, COIN creates segmentation maps that are both explanatory and functional. COIN outperforms both CAM-based attribution methods and other counterfactual methods that require full

supervision, showing that counterfactuals can be repurposed as training signals. COIN was applied to the field of medical imaging (CT scans in particular) and validated through experiments on synthetic datasets and clinical kidney tumor data from Tartu University Hospital. The counterfactual inpainting approach aims to save radiologists' time and help to debug the classifier. Using counterfactual explanations for model auditing can reveal if the classifier relies on spurious visual cues (e.g., letter markers in pneumonia detection cases), which typical attribution-based methods overlook (Ghassemi et al. 2021). Furthermore, the generated explanation must be faithful to the model. It was observed that saliency maps may work as edge detectors, misleading users' trust (Nauta et al. 2023). At the same time, the nature of counterfactual explanations in COIN provides a stronger basis for faithfulness than traditional attribution-based methods.

Nonetheless, the effectiveness of the counterfactual inpainting approach fundamentally depends on the quality and robustness of the underlying black-box classifier. Classifier imperfections such as overfitting or out-of-distribution samples may directly compromise the quality of the generated segmentation masks, and distinguishing problems arising from the classifier versus the generator remains challenging within the current architecture. Therefore, an essential avenue for future research includes integrating uncertainty estimation mechanisms into classifiers to guide counterfactual generation processes. Additionally, exploring architectures where classifiers share certain layers or parameters with the generator could further enhance both robustness and interpretability of the resulting segmentation maps.

Positioned within the broader field, COIN contributes to ongoing efforts to bridge explainability and weak supervision in computer vision, a relatively underexplored intersection. Its results suggest that counterfactuals can serve as training signals, enabling resource-efficient label generation for domains such as medical imaging, where annotation is costly and requires specialized expertise. Extending this approach to 3D modalities like MRI or volumetric CT represents a promising next step.

Trends and Opportunities for Future Research

The implications of this work contribute to the ongoing debates in the field. The interdisciplinary nature of explainability research introduces considerable confusion regarding terminology, notation, and definitions. While this slows formalization processes, it brings invaluable perspectives and a deeper understanding that ideally result in methodological innovations. We argue that the alignment with human cognitive preferences must be treated as a core design principle guiding counterfactual algorithm search. However, explanation alignment with user preferences should not be traded for faithfulness to the model. Altmeyer et al. 2024 argue that plausibility and satisfaction may come at the cost of faithfulness, which can satisfy user expectations, but misrepresent the model's logic. Additionally, Zhou et al.

2023 highlights that overly sparse explanations may act as ethical smokescreens, selectively highlighting unfair factors and concealing systemic bias, with Gilpin et al. 2018 arguing that it is fundamentally unethical. Furthermore, our research contributes to the demystification of the idea that humans generally prefer simple and sparse explanations, which is prevalent in the field, despite psychologists and cognitive scientists highlighting that it is not always the case (Hilton 1996; Keil 2006). Given the complexity of human explanation processes, it is important to consider that their causal judgments are often biased by a tendency to assign blame to abnormal factors (Kirfel et al. 2022). Therefore, generated explanations should take into account not only participants' backgrounds, but also their perceptions of normality as well as the directionality of counterfactual generation (Kuhl et al. 2023).

Considering the broader context of model evaluation, it is important to note that evaluating the satisfaction or convincingness of explanations based on user perception differs fundamentally (sometimes even orthogonally Robnik-Šikonja et al. 2018) from evaluating their factual correctness. On the other hand, disentangling whether the flaws in explanations stem from inherently flawed explanatory methods or from nonsensical relationships learned by models trained on problematic data ("garbage in, garbage out") remains challenging (Nauta et al. 2023). Given all that, it is vital to build transparent explanation methods. Otherwise, we end up building "black-box explanations" that themselves become black boxes with very little idea how to evaluate them (Rudin 2019) faithfully.

At the time of writing this thesis, the latest main language models were GPT-4.1, LLaMA 3.2, Gemini 2.0, Claude 3.5 Sonnet, Mistral Large 2, Grok-2, and DeepSeek-V3. Given recent trends and especially increasing capacities of LLMs, the field appears to be moving from static post-hoc explanations toward interactive, context-aware explanations. Systems like TalkToModel (Slack et al. 2023) produce conversational models that can send a request to explanation modules (including counterfactual explanations) and present in the form of a conversation. Furthermore, the integration of LLMs for evaluation explanations in our work, alongside concurrent research De Bona et al. 2024, brings the discussion about the evolving role of language models within human-in-the-loop. Traditionally, human-in-the-loop approaches are favored due to their potential for increasing user satisfaction and appropriately calibrated trust. However, Petsiuk et al. 2018 argues that taking humans out of the loop makes evaluation more fair to the classifier's view of the problem, therefore, better revealing the internal logic of the classifier. Given these considerations, introducing an LLM-in-the-loop paradigm could bring a promising middle ground. In this scenario, an LLM could serve as an independent oracle, assessing perceived metrics such as user satisfaction or trust. Miller 2022 argues about the necessity to distinguish between perceived trust as a subjective feeling and demonstrated trust as observable behavior, also highlighting that reliance on self-reported measures of perceived trust can obscure genuine behavioral trust outcomes, potentially leading to suboptimal calibration of trust. While per-

ceived metrics by LLMs could offer objective, scalable initial assessment, we shift our focus to demonstrated metrics (observable behaviors and practical outcomes) that cannot be effectively captured through purely automated assessments. It is worth noting that human evaluation is necessary, as LLMs may give only an averaged assessment, while every explanation is context and background dependent and should be adjusted for every recipient with the goal to fill a specific gap in their understanding (Keil 2006). Furthermore, LLM may carry training data biases or amplify existing human biases that can lead to potential harm.

In summary, the four papers presented in this thesis contribute to the improvement of algorithms, evaluation methodology, and broader adoption of counterfactual explanations. We focused on aligning explanations with human cognitive preferences on both the algorithmic and evaluation levels. Future research directions include exploring how the concepts of feasibility and coherence introduced in Chapter II can be adapted and extended to additional data modalities, such as text and images. LLM-based evaluation in Chapter III can be further refined for more nuanced user context and domain constraints to model specified experts or act as an intermediate step of LLM-in-the-loop to refine the initial assessment. At the same time, humans validate the method based on demonstrated trust in system performance. Additionally, given that we successfully modeled overall satisfaction from other explanatory metrics, a valuable next step involves deriving deeper insights to develop better quantitative evaluation metrics that more accurately capture human judgment. As for Chapter IV, a promising direction for applications is extending these techniques to handle three-dimensional medical imaging datasets such as MRI or volumetric CT. However, it would also be beneficial to expand to other applications where obtaining semantic segmentation labels is challenging and costly.

Collectively, the publications of this thesis highlight the critical importance of aligning counterfactual explanations with human cognitive preferences and demonstrate methods for evaluating and applying these explanations across various domains. We hope the present thesis contributes to valuable insights and practical tools, and eventually supports a broader adoption and increased transparency of AI-driven decisions.

BIBLIOGRAPHY

1. Abraham, Eldar D, D'Oosterlinck, Karel, Feder, Amir, Gat, Yair, Geiger, Atticus, Potts, Christopher, Reichart, Roi, and Wu, Zhengxuan (2022). "Cebab: Estimating the causal effects of real-world concepts on nlp model behavior". In: *Advances in Neural Information Processing Systems* 35, pp. 17582–17596.
2. Abrate, Carlo, Siciliano, Federico, Bonchi, Francesco, and Silvestri, Fabrizio (2024). "Human-in-the-Loop Personalized Counterfactual Recourse". In: *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland, pp. 18–38. ISBN: 978-3-031-63800-8.
3. Adadi, Amina and Berrada, Mohammed (2018). "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6, pp. 52138–52160.
4. Akula, Arjun R., Wang, Keze, Liu, Changsong, Saba-Sadiya, Sari, Lu, Hongjing, Todorovic, Sinisa, Chai, Joyce, and Zhu, Song-Chun (2022). "CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models". In: *iScience* 25.1, p. 103581. ISSN: 2589-0042. DOI: 10.1016/j.isci.2021.103581. URL: https://www.sciencedirect.com/science/article/pii/S2589004221015510.
5. Ali, Sajid, Abuhmed, Tamer, El-Sappagh, Shaker, Muhammad, Khan, Alonso-Moral, Jose M., Confalonieri, Roberto, Guidotti, Riccardo, Ser, Javier Del, Díaz-Rodríguez, Natalia, and Herrera, Francisco (2023). "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information Fusion* 99, p. 101805. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148.
6. Altmeyer, Patrick, Farmanbar, Mojtaba, Deursen, Arie van, and Liem, Cynthia CS (2024). "Faithful model explanations through energy-constrained conformal counterfactuals". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 10, pp. 10829–10837.
7. Antorán, Javier, Bhatt, Umang, Adel, Tameem, Weller, Adrian, and Hernández-Lobato, José Miguel (2020). *Getting a CLUE: A Method for Explaining Uncertainty Estimates*. DOI: 10.48550/ARXIV.2006.06848. URL: https://arxiv.org/abs/2006.06848.
8. Arrieta, Alejandro Barredo, Díaz-Rodríguez, Natalia, Del Ser, Javier, Bennetot, Adrien, Tabik, Siham, Barbado, Alberto, García, Salvador, Gil-López, Sergio, Molina, Daniel, Benjamins, Richard, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58, pp. 82–115.
9. Artelt, André and Hammer, Barbara (2019). "On the computation of counterfactual explanations–A survey". In: *arXiv preprint arXiv:1911.07749*.

10. Artelt, André and Hammer, Barbara (2020). "Convex density constraints for computing plausible counterfactual explanations". In: *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29*. Springer, pp. 353–365.
11. Artelt, André, Vaquet, Valerie, Velioglu, Riza, Hinder, Fabian, Brinkrolf, Johannes, Schilling, Malte, and Hammer, Barbara (2021). "Evaluating robustness of counterfactual explanations". In: *2021 IEEE symposium series on computational intelligence (SSCI)*. IEEE, pp. 01–09.
12. Artificial Intelligence (AI HLEG), High-Level Expert Group on (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from `https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines`.
13. Aru, Jaan, Labash, Aqeel, Corcoll, Oriol, and Vicente, Raul (2023). "Mind the gap: Challenges of deep learning approaches to theory of mind". In: *Artificial Intelligence Review* 56.9, pp. 9141–9156.
14. Arya, Vijay, Bellamy, Rachel KE, Chen, Pin-Yu, Dhurandhar, Amit, Hind, Michael, Hoffman, Samuel C, Houde, Stephanie, Liao, Q Vera, Luss, Ronny, Mojsilović, Aleksandra, et al. (2019). "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv preprint arXiv:1909.03012*.
15. Atad, Matan, Dmytrenko, Vitalii, Li, Yitong, Zhang, Xinyue, Keicher, Matthias, Kirschke, Jan, Wiestler, Bene, Khakzar, Ashkan, and Navab, Nassir (2022). "Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan". In: *arXiv preprint arXiv:2207.07553*.
16. Augustin, Maximilian, Boreiko, Valentyn, Croce, Francesco, and Hein, Matthias (2022). "Diffusion visual counterfactual explanations". In: *Advances in Neural Information Processing Systems* 35, pp. 364–377.
17. Aybek, Eren Can and Toraman, Cetin (2022). "How many response categories are sufficient for Likert type scales? An empirical study based on the Item Response Theory". In: *International Journal of Assessment Tools in Education* 9.2, pp. 534–547.
18. Bach, Sebastian, Binder, Alexander, Montavon, Grégoire, Klauschen, Frederick, Müller, Klaus-Robert, and Samek, Wojciech (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7, e0130140.
19. Bansal, Gagan, Nushi, Besmira, Kamar, Ece, Lasecki, Walter S, Weld, Daniel S, and Horvitz, Eric (2019). "Beyond accuracy: The role of mental models in human-AI team performance". In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. Vol. 7, pp. 2–11.
20. Barbu, Eduard, **Domnich**, **Marharyta**, Vicente, Raul, Sakkas, Nikos, and Morim, André (2025). "Exploring Commonalities in Explanation Frameworks: A Multi-Domain Survey Analysis". In: arXiv: `2405.11958 [cs.LG]`.

21. Becker, Barry and Kohavi, Ronny (1996). *Adult*. UCI Machine Learning Repository. "Accessed: 2024-02-18".
22. Bennett, PeterH, Burch, ThomasA, and Miller, Max (1971). "DIABETES MELLITUS IN AMERICAN (PIMA) INDIANS". In: *The Lancet* 298.7716. Originally published as Volume 2, Issue 7716, pp. 125–128. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(71)92303-8. URL: https://www.sciencedirect.com/science/article/pii/S0140673671923038.
23. Bhattacharjee, Amrita, Moraffah, Raha, Garland, Joshua, and Liu, Huan (Feb. 2024). "Towards LLM-guided Causal Explainability for Black-box Text Classifiers". In: *AAAI ReLM 2024*.
24. Bove, Clara, Lesot, Marie-Jeanne, Tijus, Charles Albert, and Detyniecki, Marcin (2023). "Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 188–203.
25. Butz, Raphaela, Hommersom, Arjen, Schulz, Renée, and Ditmarsch, Hans van (2024). "Evaluating the Usefulness of Counterfactual Explanations from Bayesian Networks". In: *Human-Centric Intelligent Systems* 4.2, pp. 286–298.
26. Byrne, Ruth M. J. (July 2019). "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, pp. 6276–6282. DOI: 10.24963/ijcai.2019/876. (Visited on 05/11/2024).
27. Byrne, Ruth MJ (2002). "Mental models and counterfactual thoughts about what might have been". In: *Trends in cognitive sciences* 6.10, pp. 426–431.
28. Byrne, Ruth MJ (2007). "Precis of the rational imagination: How people create alternatives to reality". In: *Behavioral and Brain Sciences* 30.5-6, pp. 439–453.
29. Byrne, Ruth MJ (2016). "Counterfactual thought". In: *Annual review of psychology* 67.1, pp. 135–157.
30. Byrne, Ruth MJ (2023). "Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China*, pp. 6536–6544.
31. Castelnovo, Alessandro, Depalmas, Roberto, Mercorio, Fabio, Mombelli, Nicolò, Potertı`, Daniele, Serino, Antonio, Seveso, Andrea, Sorrentino, Salvatore, and Viola, Laura (2024). "Augmenting XAI with LLMs: A Case Study in Banking Marketing Recommendation". In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 211–229.

32. Chalkidis, Ilias and Kampas, Dimitrios (2019). “Deep learning in law: early adaptation and legal word embeddings trained on large corpora”. In: *Artificial Intelligence and Law* 27.2, pp. 171–198.
33. Chang, Chun-Hao, Creager, Elliot, Goldenberg, Anna, and Duvenaud, David (2018). “Explaining image classifiers by counterfactual generation”. In: *arXiv preprint arXiv:1807.08024*.
34. Charachon, Martin, Cournede, Paul-Henry, Hudelot, Céline, and Ardon, Roberto (2022). “Leveraging conditional generative models in a general explanation framework of classifier decisions”. In: *Future Generation Computer Systems* 132, pp. 223–238.
35. Chen, Zeming, Gao, Qiyue, Bosselut, Antoine, Sabharwal, Ashish, and Richardson, Kyle (2023). *DISCO: Distilling Counterfactuals with Large Language Models*. arXiv: 2212.10534 [cs.CL]. URL: https://arxiv.org/abs/2212.10534.
36. Chen, Ziheng, Silvestri, Fabrizio, Wang, Jia, Zhu, He, Ahn, Hongshik, and Tolomei, Gabriele (2022). “Relax: Reinforcement learning agent explainer for arbitrary predictive models”. In: *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 252–261.
37. Cheng, Furui, Ming, Yao, and Qu, Huamin (2021). “DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2, pp. 1438–1447. DOI: 10.1109/TVCG.2020.3030342.
38. Chou, Yu-Liang, Moreira, Catarina, Bruza, Peter, Ouyang, Chun, and Jorge, Joaquim (2022). “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications”. In: *Information Fusion* 81, pp. 59–83.
39. Clinciu, Miruna A. and Hastie, Helen F. (Oct. 2019). “A Survey of Explainable AI Terminology”. English. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence 2019, NL4XAI 2019 ; Conference date: 29-10-2019. Association for Computational Linguistics, pp. 8–13. ISBN: 9781950737703. DOI: 10.18653/v1/W19-8403. URL: https://sites.google.com/view/nl4xai2019/.
40. Corbett-Davies, Sam, Gaebler, Johann D, Nilforoshan, Hamed, Shroff, Ravi, and Goel, Sharad (2023). “The measure and mismeasure of fairness”. In: *The Journal of Machine Learning Research* 24.1, pp. 14730–14846.
41. Covell, Jonathan (2019). *Project explAIn - Interim report*. Tech. rep. ICO: Information Commissioner’s Office (UK).
42. Dandl, Susanne, Blesch, Kristin, Freiesleben, Timo, König, Gunnar, Kapar, Jan, Bischl, Bernd, and Wright, Marvin N. (2024). “CountARFactuals – Generating

Plausible Model-Agnostic Counterfactual Explanations with Adversarial Random Forests". In: *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland, pp. 85–107. ISBN: 978-3-031-63800-8.

43. Dandl, Susanne, Molnar, Christoph, Binder, Martin, and Bischl, Bernd (2020). "Multi-objective counterfactual explanations". In: *International Conference on Parallel Problem Solving from Nature*. Springer, pp. 448–469.
44. De Bona, Francesco Bombassei, Dominici, Gabriele, Miller, Tim, Langheinrich, Marc, and Gjoreski, Martin (2024). "Evaluating explanations through llms: Beyond traditional user studies". In: *arXiv preprint arXiv:2410.17781*.
45. Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, and Zettlemoyer, Luke (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314.
46. Dhurandhar, Amit, Chen, Pin-Yu, Luss, Ronny, Tu, Chun-Chen, Ting, Paishun, Shanmugam, Karthikeyan, and Das, Payel (2018). "Explanations based on the missing: Towards contrastive explanations with pertinent negatives". In: *Advances in neural information processing systems* 31.
47. Dignum, Virginia (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 1. Springer.
48. **Domnich**, **Marharyta**, Sünter, Indrek, Trofimov, Heido, Wold, Olga, Harun, Fariha, Kostiukhin, Anton, Järveoja, Mihkel, Veske, Mihkel, Tamm, Tanel, Voormansik, Kaupo, et al. (2021). "KappaMask: AI-based cloudmask processor for Sentinel-2". In: *Remote Sensing* 13.20. ISSN: 2072-4292. DOI: 10.3390/rs13204100.
49. Domnich, Marharyta, Välja, Julius, Veski, Rasmus Moorits, Magnifico, Giacomo, Tulver, Kadi, Barbu, Eduard, and Vicente, Raul (Dec. 2024). *CounterEval: Towards Unifying Evaluation of Counterfactual Explanations*. Zenodo. DOI: 10.57967/hf/3824. URL: https://doi.org/10.57967/hf/3824.
50. **Domnich**, **Marharyta**, Välja, Julius, Veski, Rasmus Moorits, Magnifico, Giacomo, Tulver, Kadi, Barbu, Eduard, and Vicente, Raul (2025a). "Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 15, pp. 16308–16316. DOI: 10.1609/aaai.v39i15.33791.
51. **Domnich**, **Marharyta**, Veski, Rasmus Moorits, Välja, Julius, Tulver, Kadi, and Vicente, Raul (2025b). "Predicting Satisfaction of Counterfactual Explanations from Human Ratings of Explanatory Qualities". In: *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland, pp. 210–229. DOI: 10.1007/978-3-032-08317-3_10.
52. **Domnich**, **Marharyta** and Vicente, Raul (2024). "Enhancing counterfactual explanation search with diffusion distance and directional coherence". In: *Explainable Artificial Intelligence*. Vol. 2155. Communications in Computer and

Information Science. Cham: Springer Nature Switzerland, pp. 60–84. DOI: 10.1007/978-3-031-63800-8_4.

53. Doshi-Velez, Finale and Kim, Been (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.
54. Dubey, Abhimanyu, Jauhri, Abhinav, Pandey, Abhinav, Kadian, Abhishek, Al-Dahle, Ahmad, Letman, Aiesha, Mathur, Akhil, et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.
55. Ebrat, Danial and Rueda, Luis (2024). *Lusifer: LLM-based User SImulated Feedback Environment for online Recommender systems*. arXiv: 2405.13362 [cs.IR]. URL: https://arxiv.org/abs/2405.13362.
56. Ehsan, Upol, Passi, Samir, Liao, Q Vera, Chan, Larry, Lee, I-Hsiang, Muller, Michael, and Riedl, Mark O (2024). "The who in XAI: how AI background shapes perceptions of AI explanations". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–32.
57. Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, and Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon, USA: AAAI Press, pp. 226–231. DOI: 10.5555/3001460.3001507.
58. European Commission (2023). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. European Commission, Brussels. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
59. European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679: General Data Protection Regulation (GDPR)*. Official Journal of the European Union, L 119/1. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj.
60. Förster, Maximilian, Hühn, Philipp, Klier, Mathias, and Kluge, Kilian (2021). "Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations". In: *54th Hawaii International Conference on System Sciences, HICSS 2021, Kauai, Hawaii, USA, January 5, 2021*. ScholarSpace, pp. 1–10. URL: http://hdl.handle.net/10125/70767.
61. Förster, Maximilian, Hühn, Philipp, Klier, Mathias, and Kluge, Kilian (2023). "User-centric explainable AI: design and evaluation of an approach to generate coherent counterfactual explanations for structured data". In: *Journal of Decision Systems* 32.4, pp. 700–731.
62. Gao, Ge, Taymanov, Alexey, Salinas, Eduardo, Mineiro, Paul, and Misra, Dipendra (2024). *Aligning LLM Agents by Learning Latent Preference from User Edits*. arXiv: 2404.15269 [cs.CL]. URL: https://arxiv.org/abs/2404.15269.

63. Ge, Yingqiang, Tan, Juntao, Zhu, Yan, Xia, Yinglong, Luo, Jiebo, Liu, Shuchang, Fu, Zuohui, Geng, Shijie, Li, Zelong, and Zhang, Yongfeng (2022). "Explainable fairness in recommendation". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 681–691.
64. Gentile, Claudio and Warmuth, Manfred KK (1998). "Linear hinge loss and average margin". In: *Advances in neural information processing systems* 11.
65. Ghassemi, Marzyeh, Oakden-Rayner, Luke, and Beam, Andrew L (2021). "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11, e745–e750.
66. Ghazimatin, Azin, Balalau, Oana, Saha Roy, Rishiraj, and Weikum, Gerhard (2020). "Prince: Provider-side interpretability with counterfactual explanations in recommender systems". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 196–204.
67. Gilpin, Leilani H, Bau, David, Yuan, Ben Z, Bajwa, Ayesha, Specter, Michael, and Kagal, Lalana (2018). "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, pp. 80–89.
68. Girotra, Karan, Meincke, Lennart, Terwiesch, Christian, and Ulrich, Karl T. (2023). "Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation". In: *SSRN Electronic Journal*. URL: `https://api.semanticscholar.org/CorpusID:260467886`.
69. Goethals, Sofie, Martens, David, and Calders, Toon (2024). "PreCoF: counterfactual explanations for fairness". In: *Machine Learning* 113.5, pp. 3111–3142.
70. Google (2023). *Explainable AI tools*. Google Cloud. Retrieved from `https://cloud.google.com/explainable-ai`.
71. Goyal, Yash, Wu, Ziyan, Ernst, Jan, Batra, Dhruv, Parikh, Devi, and Lee, Stefan (2019). "Counterfactual visual explanations". In: *International Conference on Machine Learning*. PMLR, pp. 2376–2384.
72. Graziani, Mara, Dutkiewicz, Lidia, Calvaresi, Davide, Amorim, José Pereira, Yordanova, Katerina, Vered, Mor, Nair, Rahul, Abreu, Pedro Henriques, Blanke, Tobias, Pulignano, Valeria, et al. (2023). "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences". In: *Artificial intelligence review* 56.4, pp. 3473–3504.
73. Guidotti, Riccardo (Apr. 28, 2022). "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: `10.1007/s10618-022-00831-6`. URL: `https://doi.org/10.1007/s10618-022-00831-6`.
74. Gunning, David and Aha, David (2019). "DARPA's explainable artificial intelligence (XAI) program". In: *AI magazine* 40.2, pp. 44–58.

75. Guo, Hangzhi, Jia, Feiran, Chen, Jinghui, Squicciarini, Anna, and Yadav, Amulya (2023). "RoCourseNet: Robust Training of a Prediction Aware Recourse Model". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. Birmingham, United Kingdom: Association for Computing Machinery, pp. 619–628. ISBN: 9798400701245. DOI: 10.1145/3583780.3615040. URL: https://doi.org/10.1145/3583780.3615040.
76. Heaton, James B, Polson, Nick G, and Witte, Jan Hendrik (2017). "Deep learning for finance: deep portfolios". In: *Applied Stochastic Models in Business and Industry* 33.1, pp. 3–12.
77. Hilton, Denis J (1996). "Mental models and causal explanation: Judgements of probable cause and explanatory relevance". In: *Thinking & Reasoning* 2.4, pp. 273–308.
78. Hoffman, Robert R, Mueller, Shane T, Klein, Gary, and Litman, Jordan (2018). "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608*.
79. Hoffman, Robert R, Mueller, Shane T, Klein, Gary, and Litman, Jordan (2023). "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance". In: *Frontiers in Computer Science* 5, p. 1096257.
80. Hofmann, Hans (1994). *Statlog (German Credit Data)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.
81. Holzinger, Andreas, Saranti, Anna, Angerschmid, Alessa, Finzel, Bettina, Schmid, Ute, and Mueller, Heimo (2023). "Toward human-level concept learning: Pattern benchmarking for AI algorithms". In: *Patterns*.
82. Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, and Chen, Weizhu (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685. (Visited on 04/13/2024).
83. Jeanneret, Guillaume, Simon, Loïc, and Jurie, Frédéric (2022). "Diffusion models for counterfactual explanations". In: *Proceedings of the Asian conference on computer vision*, pp. 858–876.
84. Jiang, Junqi, Leofante, Francesco, Rago, Antonio, and Toni, Francesca (2024). "Robust counterfactual explanations in machine learning: A survey". In: *arXiv preprint arXiv:2402.01928*.
85. Jin, Ming, Wang, Shiyu, Ma, Lintao, Chu, Zhixuan, Zhang, James Y., Shi, Xiaoming, Chen, Pin-Yu, Liang, Yuxuan, Li, Yuan-Fang, Pan, Shirui, and Wen, Qingsong (2024). *Time-LLM: Time Series Forecasting by Reprogramming Large Language Models*. arXiv: 2310.01728 [cs.LG]. URL: https://arxiv.org/abs/2310.01728.
86. Johnson-Laird, Philip N (2010). "Mental models and human reasoning". In: *Proceedings of the National Academy of Sciences* 107.43, pp. 18243–18250.

87. Kahneman, Daniel (2011). “Fast and slow thinking”. In: *Allen Lane and Penguin Books, New York*.
88. Kanamori, Kentaro, Takagi, Takuya, Kobayashi, Ken, and Arimura, Hiroki (2020). “DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization.” In: *IJCAI*, pp. 2855–2862.
89. Karimi, Amir-Hossein, Barthe, Gilles, Schölkopf, Bernhard, and Valera, Isabel (2022). “A survey of algorithmic recourse: contrastive explanations and consequential recommendations”. In: *ACM Computing Surveys* 55.5, pp. 1–29.
90. Karimi, Amir-Hossein, Schölkopf, Bernhard, and Valera, Isabel (2021). “Algorithmic recourse: from counterfactual explanations to interventions”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362.
91. Kaushik, Divyansh, Hovy, Eduard, and Lipton, Zachary C (2019). “Learning the difference that makes a difference with counterfactually-augmented data”. In: *arXiv preprint arXiv:1909.12434*.
92. Keane, Mark T, Kenny, Eoin M, Delaney, Eoin, and Smyth, Barry (2021a). *If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques*. arXiv: 2103.01035 [cs.LG]. URL: https://arxiv.org/abs/2103.01035.
93. Keane, Mark T, Kenny, Eoin M, Delaney, Eoin, and Smyth, Barry (2021b). “If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques”. In: *arXiv preprint arXiv:2103.01035*.
94. Keil, Frank C. (2006). “Explanation and understanding.” In: *Annual review of psychology* 57, pp. 227–254. ISSN: 0066-4308 1545-2085. DOI: 10.1146/annurev.psych.57.102904.190100.
95. Kenny, Eoin M, Ford, Courtney, Quinn, Molly, and Keane, Mark T (2021). “Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies”. In: *Artificial Intelligence* 294, p. 103459.
96. Khajuria, Tarun, Dias, Braian Olmiro, **Domnich**, **Marharyta**, and Aru, Jaan (2025). “Interpreting the structure of multi-object representations in vision encoders”. In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 359–382. DOI: 10.1007/978-3-032-08324-1_16.
97. Kim, Hyunwoo, Choi, Yoonseo, Yang, Taehyun, Lee, Honggu, Park, Chaneon, Lee, Yongju, Kim, Jin Young, and Kim, Juho (2024). *Using LLMs to Investigate Correlations of Conversational Follow-up Queries with User Satisfaction*. arXiv: 2407.13166 [cs.HC]. URL: https://arxiv.org/abs/2407.13166.
98. Kirfel, Lara, Icard, Thomas, and Gerstenberg, Tobias (2022). “Inference from explanation.” In: *Journal of Experimental Psychology: General* 151.7, p. 1481.

99. Kirsch, Alexandra (2017). “Explain to whom? Putting the user in the center of explainable AI”. In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017)*.
100. Kiseleva, Julia, Williams, Kyle, Hassan Awadallah, Ahmed, Crook, Aidan C., Zitouni, Imed, and Anastasakos, Tasos (2016). “Predicting User Satisfaction with Intelligent Assistants”. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’16. Pisa, Italy: Association for Computing Machinery, pp. 45–54. ISBN: 9781450340694. DOI: 10.1145/2911451.2911521. URL: https://doi.org/10.1145/2911451.2911521.
101. Klein, Lukas, El-Assady, Mennatallah, and Jäger, Paul F. (2022). *From Correlation to Causation: Formalizing Interpretable Machine Learning as a Statistical Process*. arXiv: 2207.04969 [cs.CV].
102. Kliegr, Tomáš, Bahník, Štěpán, and Fürnkranz, Johannes (2021). “A review of possible effects of cognitive biases on interpretation of rule-based machine learning models”. In: *Artificial Intelligence* 295, p. 103458.
103. Krosnick, Jon A (1991). “Response strategies for coping with the cognitive demands of attitude measures in surveys”. In: *Applied cognitive psychology* 5.3, pp. 213–236.
104. Kshetry, Neelabh and Kantardzic, Mehmed (2024). “What-if XAI framework (WiXAI): from counterfactuals towards causal understanding”. In: *Journal of Computer and Communications* 12.6, pp. 169–198.
105. Kuhl, Ulrike, Artelt, André, and Hammer, Barbara (2023). “For Better or Worse: The Impact of Counterfactual Explanations’ Directionality on User Behavior in xAI”. In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 280–300.
106. Kulesza, Todd, Stumpf, Simone, Burnett, Margaret, Yang, Sherry, Kwan, Irwin, and Wong, Weng-Keen (2013). “Too much, too little, or just right? Ways explanations impact end users’ mental models”. In: *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, pp. 3–10.
107. Laugel, Thibault, Lesot, Marie-Jeanne, Marsala, Christophe, and Detyniecki, Marcin (2019a). “Issues with post-hoc counterfactual explanations: a discussion”. In: *arXiv preprint arXiv:1906.04774*.
108. Laugel, Thibault, Lesot, Marie-Jeanne, Marsala, Christophe, Renard, Xavier, and Detyniecki, Marcin (2018). “Comparison-based inverse classification for interpretability in machine learning”. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations: 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part I 17*. Springer, pp. 100–111.

109. Laugel, Thibault, Lesot, Marie-Jeanne, Marsala, Christophe, Renard, Xavier, and Detyniecki, Marcin (2019b). *The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations*. arXiv: 1907.09294 [cs.LG]. URL: https://arxiv.org/abs/1907.09294.
110. Le, Thao, Miller, Tim, Singh, Ronal, and Sonenberg, Liz (2022). "Improving Model Understanding and Trust with Counterfactual Explanations of Model Confidence". In: *arXiv preprint arXiv:2206.02790*.
111. Leofante, Francesco and Potyka, Nico (2024). "Promoting counterfactual robustness through diversity". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 19, pp. 21322–21330.
112. Lewis, David (1973). "Counterfactuals and comparative possibility". In: *IFS: Conditionals, Belief, Decision, Chance and Time*. Springer, pp. 57–85.
113. Ley, Dan, Bhatt, Umang, and Weller, Adrian (June 2022). "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.7, pp. 7390–7398. DOI: 10.1609/aaai.v36i7.20702. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20702.
114. Lipton, Zachary C (2018). "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.
115. Liu, Yang, Khandagale, Sujay, White, Colin, and Neiswanger, Willie (2021). "Synthetic benchmarks for scientific research in explainable machine learning". In: *arXiv preprint arXiv:2106.12543*.
116. Liu, Yang, Yao, Yuanshun, Ton, Jean-Francois, Zhang, Xiaoying, Guo, Ruocheng, Cheng, Hao, Klochkov, Yegor, Taufiq, Muhammad Faaiz, and Li, Hang (2024). *Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment*. arXiv: 2308.05374 [cs.AI]. URL: https://arxiv.org/abs/2308.05374.
117. Lombrozo, Tania (2007). "Simplicity and probability in causal explanation". In: *Cognitive psychology* 55.3, pp. 232–257.
118. Longo, Luca, Brcic, Mario, Cabitza, Federico, Choi, Jaesik, Confalonieri, Roberto, Del Ser, Javier, Guidotti, Riccardo, Hayashi, Yoichi, Herrera, Francisco, Holzinger, Andreas, et al. (2024). "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions". In: *Information Fusion* 106, p. 102301.
119. Lucic, Ana, Oosterhuis, Harrie, Haned, Hinda, and Rijke, Maarten de (2022). "FOCUS: Flexible optimizable counterfactual explanations for tree ensembles". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 5, pp. 5313–5322.
120. Lundberg, Scott M and Lee, Su-In (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

121. Maaten, Laurens van der and Hinton, Geoffrey (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.
122. Mackonis, Adolfas (2013). “Inference to the best explanation, coherence and other explanatory virtues”. In: *Synthese* 190.6, pp. 975–995.
123. Mahajan, Divyat, Tan, Chenhao, and Sharma, Amit (2019). “Preserving causal constraints in counterfactual explanations for machine learning classifiers”. In: *arXiv preprint arXiv:1912.03277*.
124. Majoral, Daniel and **Domnich**, **Marharyta** (May 2025). “Kaizen: Decomposing cellular images with VQ-VAE”. In: *PLOS ONE* 20.5, pp. 1–11. DOI: 10.1371/journal.pone.0313549.
125. McCloy, Rachel and Byrne, Ruth MJ (2000). “Counterfactual thinking about controllable events”. In: *Memory & Cognition* 28, pp. 1071–1078.
126. Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina, and Galstyan, Aram (2021). “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6, pp. 1–35.
127. Mertes, Silvan, Huber, Tobias, Weitz, Katharina, Heimerl, Alexander, and André, Elisabeth (2022). “Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning”. In: *Frontiers in artificial intelligence* 5, p. 825565.
128. Miller, Tim (2019). “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial Intelligence* 267, pp. 1–38. ISSN: 0004-3702. DOI: 10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.
129. Miller, Tim (2021). “Contrastive explanation: a structural-model approach”. In: *The Knowledge Engineering Review* 36. Edition: 2021/10/20 Publisher: Cambridge University Press, e14. ISSN: 0269-8889. DOI: 10.1017/S0269888921000102. URL: https://www.cambridge.org/core/article/contrastive-explanation-a-structuralmodel-approach/69A2E32B160C2C7FB65BC88670D7AEA7.
130. Miller, Tim (2022). “Are we measuring trust correctly in explainability, interpretability, and transparency research?” In: *arXiv preprint arXiv:2209.00651*.
131. Miotto, Riccardo, Wang, Fei, Wang, Shuang, Jiang, Xiaoqian, and Dudley, Joel T (2018). “Deep learning for healthcare: review, opportunities and challenges”. In: *Briefings in bioinformatics* 19.6, pp. 1236–1246.
132. Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.
133. Moreira, Catarina, Chou, Yu-Liang, Hsieh, Chihcheng, Ouyang, Chun, Jorge, Joaquim, and Pereira, João Madeiras (2022). “Benchmarking Counterfactual Algorithms for XAI: From White Box to Black Box”. In: *arXiv preprint arXiv:2203.02399*.

134. Mothilal, Ramaravind K., Sharma, Amit, and Tan, Chenhao (2020). "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. event-place: Barcelona, Spain. New York, NY, USA: Association for Computing Machinery, pp. 607–617. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372850. URL: https://doi.org/10.1145/3351095.3372850.

135. Mueller, Shane T, Hoffman, Robert R, Clancey, William, Emrey, Abigail, and Klein, Gary (2019). "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI". In: *arXiv preprint arXiv:1902.01876*.

136. Nauta, Meike, Trienes, Jan, Pathak, Shreyasi, Nguyen, Elisa, Peters, Michelle, Schmitt, Yasmin, Schlötterer, Jörg, Van Keulen, Maurice, and Seifert, Christin (2023). "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai". In: *ACM Computing Surveys* 55.13s, pp. 1–42.

137. Nemirovsky, Daniel, Thiebaut, Nicolas, Xu, Ye, and Gupta, Abhishek (2022). "Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 1488–1497.

138. Nori, Harsha, Jenkins, Samuel, Koch, Paul, and Caruana, Rich (2019). "Interpretml: A unified framework for machine learning interpretability". In: *arXiv preprint arXiv:1909.09223*.

139. Obermeyer, Ziad, Powers, Brian, Vogeli, Christine, and Mullainathan, Sendhil (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464, pp. 447–453.

140. OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774.

141. Parmentier, Axel and Vidal, Thibaut (2021). "Optimal counterfactual explanations in tree ensembles". In: *International conference on machine learning*. PMLR, pp. 8422–8431.

142. Pawelczyk, Martin, Bielawski, Sascha, Heuvel, Johannes van den, Richter, Tobias, and Kasneci, Gjergji (2021). *CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms*. DOI: 10.48550/ARXIV.2108.00783. URL: https://arxiv.org/abs/2108.00783.

143. Pearl, Judea (2009). *Causality*. Cambridge university press.

144. Pearl, Judea (2019). "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3, pp. 54–60.

145. Perrig, Sebastian AC, Scharowski, Nicolas, and Brühlmann, Florian (2023). "Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI". In: *Extended abstracts of the 2023 CHI Conference on human factors in computing systems*, pp. 1–7.

146. Petsiuk, Vitali, Das, Abir, and Saenko, Kate (2018). “Rise: Randomized input sampling for explanation of black-box models”. In: *arXiv preprint arXiv:1806.07421*.
147. Poché, Antonin, Hervier, Lucas, and Bakkay, Mohamed-Chafik (2023). “Natural example-based explainability: a survey”. In: *World Conference on eXplainable Artificial Intelligence*. Springer, pp. 24–47.
148. Poyiadzi, Rafael, Sokol, Kacper, Santos-Rodriguez, Raul, Bie, Tijl De, and Flach, Peter (2020). “FACE”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM. DOI: 10.1145/3375627.3375850. URL: https://doi.org/10.1145%2F3375627.3375850.
149. Prasad, NG Narasimha and Rao, Jon NK (1990). “The estimation of the mean squared error of small-area estimators”. In: *Journal of the American statistical association* 85.409, pp. 163–171.
150. Rajani, Nazneen Fatema, McCann, Bryan, Xiong, Caiming, and Socher, Richard (2019). “Explain yourself! leveraging language models for commonsense reasoning”. In: *arXiv preprint arXiv:1906.02361*.
151. Raman, Natraj, Magazzeni, Daniele, and Shah, Sameena (2023). “Bayesian hierarchical models for counterfactual estimation”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1115–1128.
152. Ramon, Yanou, Vermeire, Tom, Toubia, Olivier, Martens, David, and Evgeniou, Theodoros (2021). “Understanding consumer preferences for explanations generated by XAI algorithms”. In: *arXiv preprint arXiv:2107.02624*.
153. Rasouli, Peyman and Chieh Yu, Ingrid (2024). “CARE: Coherent actionable recourse based on sound counterfactual explanations”. In: *International Journal of Data Science and Analytics* 17.1, pp. 13–38.
154. Reddy, Chandan KA, Gopal, Vishak, and Cutler, Ross (2022). “DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 886–890.
155. Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos (2016). “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
156. Robnik-Šikonja, Marko and Bohanec, Marko (2018). “Perturbation-based explanations of prediction models”. In: *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175.
157. Rosenfeld, Avi (2021). “Better metrics for evaluating explainable artificial intelligence”. In: *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pp. 45–50.

158. Rudin, Cynthia (2019). “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5, pp. 206–215.
159. Russell, Chris (2019). “Efficient search for diverse coherent explanations”. In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 20–28.
160. Sakkas, Nikos, Yfanti, Sofia, Daskalakis, Costas, Barbu, Eduard, and **Domnich**, **Marharyta** (2021). “Interpretable forecasting of energy demand in the residential sector”. In: *Energies* 14.20. ISSN: 1996-1073. DOI: 10.3390/en14206568.
161. Sakkas, Nikos, Yfanti, Sofia, Shah, Pooja, Sakkas, Nikitas, Chaniotakis, Christina, Daskalakis, Costas, Barbu, Eduard, and **Domnich**, **Marharyta** (2023). “Explainable approaches for forecasting building electricity consumption”. In: *Energies* 16.20. ISSN: 1996-1073. DOI: 10.3390/en16207210.
162. Salewski, Leonard, Koepke, A Sophia, Lensch, Hendrik PA, and Akata, Zeynep (2020). “Clevr-x: A visual reasoning dataset for natural language explanations”. In: *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, pp. 69–88.
163. Sapienza, Alessandro, Cantucci, Filippo, and Falcone, Rino (2022). “Modeling Interaction in Human–Machine Systems: A Trust and Trustworthiness Approach”. In: *Automation* 3.2, pp. 242–257. ISSN: 2673-4052. DOI: 10.3390/automation3020012. URL: https://www.mdpi.com/2673-4052/3/2/12.
164. Scharowski, Nicolas, Perrig, Sebastian AC, Aeschbach, Lena Fanya, Felten, Nick von, Opwis, Klaus, Wintersberger, Philipp, and Brühlmann, Florian (2024). “To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI”. In: *arXiv preprint arXiv:2403.00582*.
165. Schmidt, Philipp, Biessmann, Felix, and Teubner, Timm (2020). “Transparency and trust in artificial intelligence systems”. In: *Journal of Decision Systems* 29.4, pp. 260–278.
166. Schwalbe, Gesina and Finzel, Bettina (2024). “A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts”. In: *Data Mining and Knowledge Discovery* 38.5, pp. 3043–3101.
167. Selvaraju, Ramprasaath R, Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Parikh, Devi, and Batra, Dhruv (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.
168. Sharma, Shubham, Henderson, Jette, and Ghosh, Joydeep (2019). “Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models”. In: *arXiv preprint arXiv:1905.07857*.

169. Shen, Dinggang, Wu, Guorong, and Suk, Heung-Il (2017). "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19, pp. 221–248.
170. Shrikumar, Avanti, Greenside, Peyton, and Kundaje, Anshul (2017). "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMlR, pp. 3145–3153.
171. Shvetsov, Dmytro, Ariva, Joonas, **Domnich**, **Marharyta**, Vicente, Raul, and Fishman, Dmytro (2024). "COIN: Counterfactual inpainting for weakly supervised semantic segmentation for medical images". In: *Explainable Artificial Intelligence*. Vol. 2155. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, pp. 45–59. DOI: 10.1007/978-3-031-63800-8_3.
172. Singla, Sumedha, Eslami, Motahhare, Pollack, Brian, Wallace, Stephen, and Batmanghelich, Kayhan (2023). "Explaining the black-box smoothly—a counterfactual approach". In: *Medical Image Analysis* 84, p. 102721.
173. Siro, Clemencia, Aliannejadi, Mohammad, and De Rijke, Maarten (Nov. 2023). "Understanding and Predicting User Satisfaction with Conversational Recommender Systems". In: *ACM Trans. Inf. Syst.* 42.2. ISSN: 1046-8188. DOI: 10.1145/3624989. URL: https://doi.org/10.1145/3624989.
174. Sithakoul, Samuel, Meftah, Sara, and Feutry, Clément (2024). "Beexai: Benchmark to evaluate explainable ai". In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 445–468.
175. Slack, Dylan, Hilgard, Anna, Lakkaraju, Himabindu, and Singh, Sameer (2021a). "Counterfactual explanations can be manipulated". In: *Advances in neural information processing systems* 34, pp. 62–75.
176. Slack, Dylan, Hilgard, Anna, Singh, Sameer, and Lakkaraju, Himabindu (2021b). "Reliable post hoc explanations: Modeling uncertainty in explainability". In: *Advances in neural information processing systems* 34, pp. 9391–9404.
177. Slack, Dylan, Krishna, Satyapriya, Lakkaraju, Himabindu, and Singh, Sameer (July 2023). "Explaining machine learning models with interactive natural language conversations using TalkToModel". In: *Nature Machine Intelligence*. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00692-8. URL: https://doi.org/10.1038/s42256-023-00692-8.
178. Sokol, Kacper and Flach, Peter (2020). "Explainability fact sheets: A framework for systematic assessment of explainable approaches". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 56–67.
179. Spreitzer, Nina, Haned, Hinda, and Linden, Ilse van der (2022). "Evaluating the Practicality of Counterfactual Explanations." In: *XAI. it@ AI* IA*, pp. 31–50.
180. Stepin, Ilia, Alonso, Jose M., Catala, Alejandro, and Pereira-Fariña, Martín (2021). "A Survey of Contrastive and Counterfactual Explanation Generation

Methods for Explainable Artificial Intelligence". In: *IEEE Access* 9, pp. 11974–12001. DOI: 10.1109/ACCESS.2021.3051315.

181. Stepin, Ilia, Alonso-Moral, Jose M., Catala, Alejandro, and Pereira-Fariña, Martín (2022). "An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information". In: *Information Sciences* 618, pp. 379–399. ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.10.098. URL: https://www.sciencedirect.com/science/article/pii/S002002552201218X.
182. Strickland, Brent and Keil, Frank (2011). "Event completion: Event based inferences distort memory in a matter of seconds". In: *Cognition* 121.3, pp. 409–415.
183. Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.
184. Templeton, Adly, Conerly, Thomas, Marcus, Jonathan, Lindsey, Jack, Bricken, Trenton, Chen, Brian, Pearce, Adam, Citro, Craig, Ameisen, Emmanuel, Jones, Andy, Cunningham, Hoagy, Turner, Nicholas L., McDougall, Callum, MacDiarmid, Monte, Freeman, C. Daniel, and Sumers, Theodore R. (2024). *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Transformer Circuits Thread. URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/.
185. Tešić, Marko and Hahn, Ulrike (2022). "Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that?" In: *Patterns* 3.12.
186. Thagard, Paul (1989). "Explanatory Coherence (Plus Commentary)". In: *Behavioral and Brain Sciences* 12.3, pp. 435–467. DOI: 10.1017/s0140525x00057046.
187. The White House (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. The White House, Washington D.C. Retrieved from https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
188. Thommes, Kirsten, Lammert, Olesja, Schütze, Christian, Richter, Birte, and Wrede, Britta (2024). "Human emotions in AI explanations". In: *World Conference on Explainable Artificial Intelligence*. Springer, pp. 270–293.
189. Tversky, Amos and Simonson, Itamar (1993). "Context-dependent preferences". In: *Management science* 39.10, pp. 1179–1189.
190. University of Tartu (2018). *UT Rocket*. DOI: 10.23673/PH6N-0144.

191. Upadhyay, Sohini, Joshi, Shalmali, and Lakkaraju, Himabindu (2021). "Towards robust and reliable algorithmic recourse". In: *Advances in Neural Information Processing Systems* 34, pp. 16926–16937.
192. Valja, Julius (2024). *Assessing the Quality of Counterfactual Explanations with Large Language Models*. University of Tartu, Institute of Computer Science, Bachelor's thesis.
193. Van Looveren, Arnaud and Klaise, Janis (2021a). "Interpretable counterfactual explanations guided by prototypes". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 650–665.
194. Van Looveren, Arnaud, Klaise, Janis, Vacanti, Giovanni, and Cobb, Oliver (2021b). "Conditional generative models for counterfactual explanations". In: *arXiv preprint arXiv:2101.10123*.
195. Vandenhende, Simon, Mahajan, Dhruv, Radenovic, Filip, and Ghadiyaram, Deepti (2022). "Making heads or tails: Towards semantically consistent visual counterfactuals". In: *European Conference on Computer Vision*. Springer, pp. 261–279.
196. VanNostrand, Peter M, Hofmann, Dennis M, Ma, Lei, and Rundensteiner, Elke A (2024). "Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1682–1700.
197. Verma, Sahil, Boonsanong, Varich, Hoang, Minh, Hines, Keegan E, Dickerson, John P, and Shah, Chirag (2020). "Counterfactual explanations and algorithmic recourses for machine learning: A review". In: *arXiv preprint arXiv:2010.10596*.
198. Vilone, Giulia and Longo, Luca (2021). "Notions of explainability and evaluation approaches for explainable artificial intelligence". In: *Information Fusion* 76, pp. 89–106. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.05.009. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001093.
199. Vilone, Giulia and Longo, Luca (2023). "Development of a human-centred psychometric test for the evaluation of explanations produced by XAI methods". In: *World Conference on Explainable Artificial Intelligence*. Springer Nature Switzerland Cham, pp. 205–232.
200. Virgolin, Marco, Alderliesten, Tanja, Witteveen, Cees, and Bosman, Peter A. N. (2021). "Improving model-based genetic programming for symbolic regression of small expressions". In: *Evolutionary Computation* 29.2, pp. 211–237.
201. Virgolin, Marco and Fracaros, Saverio (2023). "On the robustness of sparse counterfactual explanations to adverse perturbations". In: *Artificial Intelligence* 316, p. 103840.

202. Wachter, Sandra, Mittelstadt, Brent, and Russell, Chris (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31, p. 841.

203. Wang, Haofan, Wang, Zifan, Du, Mengnan, Yang, Fan, Zhang, Zijian, Ding, Sirui, Mardziel, Piotr, and Hu, Xia (2020). "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25.

204. Wang, Lei, Ma, Chen, Feng, Xueyang, Zhang, Zeyu, Yang, Hao, Zhang, Jingsen, Chen, Zhiyuan, Tang, Jiakai, Chen, Xu, Lin, Yankai, Zhao, Wayne Xin, Wei, Zhewei, and Wen, Jirong (Mar. 2024). "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6, p. 186345. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: https://doi.org/10.1007/s11704-024-40231-1.

205. Wang, Xiangmeng, Li, Qian, Yu, Dianer, Li, Qing, and Xu, Guandong (2024). "Counterfactual explanation for fairness in recommendation". In: *ACM Transactions on Information Systems* 42.4, pp. 1–30.

206. Wang, Xinru and Yin, Ming (2021). "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making". In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 318–328.

207. Wang, Yongjie, Qian, Hangwei, Liu, Yongjie, Guo, Wei, and Miao, Chunyan (2023). "Flexible and robust counterfactual explanations with minimal satisfiable perturbations". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2596–2605.

208. Wang, Zhenhailong, Mao, Shaoguang, Wu, Wenshan, Ge, Tao, Wei, Furu, and Ji, Heng (June 2024). "Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 257–279. DOI: 10.18653/v1/2024.naacl-long.15. URL: https://aclanthology.org/2024.naacl-long.15.

209. Warren, Greta, Byrne, Ruth M. J., and Keane, Mark T. (2023). "Categorical and Continuous Features in Counterfactual Explanations of AI Systems". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. Sydney, NSW, Australia: Association for Computing Machinery, pp. 171–187. ISBN: 9798400701061. DOI: 10.1145/3581641.3584090. URL: https://doi.org/10.1145/3581641.3584090.

210. Warren, Greta, Keane, Mark T, and Byrne, Ruth MJ (2022). "Features of Explainability: How users understand counterfactual and causal explana-

tions for categorical and continuous features in XAI". In: *arXiv preprint arXiv:2204.10152*.

211. Wei, Yinwei, Qu, Xiaoyang, Wang, Xiang, Ma, Yunshan, Nie, Liqiang, and Chua, Tat-Seng (2023). "Rule-guided Counterfactual Explainable Recommendation". In: *IEEE Transactions on Knowledge and Data Engineering*.
212. Whittaker, Meredith, Crawford, Kate, Dobbe, Roel, Fried, Genevieve, Kaziunas, Elizabeth, Mathur, Varoon, West, Sarah Mysers, Richardson, Rashida, Schultz, Jason, Schwartz, Oscar, et al. (2018). *AI now report 2018*. AI Now Institute at New York University New York.
213. Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W. (1995). *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.
214. Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Perric, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Le Scao, Teven, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, and Rush, Alexander M. (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. (Visited on 04/15/2024).
215. Wu, Haochen, Sharma, Shubham, Patra, Sunandita, and Gopalakrishnan, Sriram (Mar. 2024). "SafeAR: Safe Algorithmic Recourse by Risk-Aware Policies". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.14, pp. 15915–15923. ISSN: 2159-5399. DOI: 10.1609/aaai.v38i14.29522. URL: http://dx.doi.org/10.1609/aaai.v38i14.29522.
216. Yang, Chengrun, Wang, Xuezhi, Lu, Yifeng, Liu, Hanxiao, Le, Quoc V., Zhou, Denny, and Chen, Xinyun (2024). *Large Language Models as Optimizers*. arXiv: 2309.03409 [cs.LG]. URL: https://arxiv.org/abs/2309.03409.
217. Yang, Zhaojun, Levow, Gina-Anne, and Meng, Helen (2012). "Predicting User Satisfaction in Spoken Dialog System Evaluation With Collaborative Filtering". In: *IEEE Journal of Selected Topics in Signal Processing* 6.8, pp. 971–981. DOI: 10.1109/JSTSP.2012.2229965.
218. You, Dianlong, Niu, Shina, Dong, Siqi, Yan, Huigui, Chen, Zhen, Wu, Di, Shen, Limin, and Wu, Xindong (2023). "Counterfactual explanation generation with minimal feature boundary". In: *Information Sciences* 625, pp. 342–366.
219. Yue, Zhiling, Fang, Yingying, Yang, Liutao, Baid, Nikhil, Walsh, Simon, and Yang, Guang (2024). "Enhancing Weakly Supervised Semantic Segmentation for Fibrosis via Controllable Image Generation". In: *arXiv preprint arXiv:2411.03551*.
220. Zelnik-Manor, Lihi and Perona, Pietro (2004). "Self-tuning spectral clustering". In: *Advances in neural information processing systems* 17.

221. Zemla, Jeffrey C, Sloman, Steven, Bechlivanidis, Christos, and Lagnado, David A (2017). “Evaluating everyday explanations”. In: *Psychonomic bulletin & review* 24, pp. 1488–1500.
222. Zemni, Mehdi, Chen, Mickaël, Zablocki, Éloi, Ben-Younes, Hédi, Pérez, Patrick, and Cord, Matthieu (2023). “Octet: Object-aware counterfactual explanations”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15062–15071.
223. Zhou, Joyce and Joachims, Thorsten (2023). “How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making”. In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 12–21.

ACKNOWLEDGEMENTS

First of all, I would like to thank everyone who contributed to scaling and developing the Institute of Computer Science at the University of Tartu. It may seem like a coincidence that life brought me to Estonia and that I pursued my PhD here, but in reality, it is the result of the work of many people who built the Institute's reputation. Honorable mention goes to Prof. Jaak Vilo for his leadership, and to Dmytro Fishman, whose enthusiastic promotion of Tartu in Ukraine during my bachelor's studies encouraged me to apply for a master's studies here. These events and receiving a scholarship ultimately shaped the path that brought me to this point. The excellent teaching I received during my master's studies from Dmytro Fishman and Mari-Liis Allikivi also played an important role, as it showed me how inspiring and approachable PhD students can be. Seeing Mari-Liis as a researcher and teacher at the Institute gave me a female role model to look up to and a path I could imagine following.

After completing my master's degree, I decided to pursue a PhD to stay in a community of brilliant people passionate about knowledge and science. I am happy to report that the people I met during these years exceeded all expectations.

Supervisors. I would like to express my deepest gratitude to my supervisor, Prof. Raul Vicente, for his guidance, trust, and patience throughout this journey. Thank you for our many conversations and for allowing and enabling me to pursue all of my ideas in parallel, helping to prioritize and watching my back when things were colliding. My PhD path was far from straightforward. I was part of the EU project TRUST-AI, where, as key partners, we were expected to coordinate, collaborate, and deliver a wide range of results. At the same time, I constantly found myself starting new side projects and collaborations every other month. I am grateful for Raul's trust and for the sense of support that allowed me to explore so freely. I still remember the moment when I decided to submit three papers to the same deadline (I think I saw doubts and fear in his eyes), but I was never discouraged from doing what I believed was right. That freedom is a true privilege. Thank you for your continuous support.

I would also like to thank my co-supervisor, Eduard Barbu, for shielding me from additional side projects in the later stages of my PhD and for taking care of monthly meetings when I was stretched too thin.

Special thanks to Jaan Aru for his kindness and for the thoughtful, delicate words of support during difficult times. I am grateful to both Raul and Jaan for building NAIL (Computational Neuroscience in the past) lab full of diverse, smart, and genuinely kind people.

Reviewers. I would like to thank my internal and external reviewers. Prof. Mark Fišel, thank you for your constructive feedback and for bringing humor to your comments. I definitely did not expect to laugh while reviewing my corrections. Big thanks to my external reviewer Prof. Barbara Hammer, your words of praise were an exceptionally rare gift in academia, and I will treasure them for the

rest of my career. Prof. Luca Longo, thank you for your thoughtful review and also for your tremendous work in building the Explainable AI community. Attending the first World Conference on eXplainable Artificial Intelligence (xAI 2022) in Portugal was a turning point for me personally. It was the first time I truly felt a sense of belonging in academia. That conference was the right size, with the right people, and it has since become my annual tradition to present there. The event provides an incredible sense of community and connection with leading researchers in XAI. Over the years, I have seen both young scientists and this young field grow together. Thank you, Luca, for making this possible and shaping XAI community!

Collaborators. During my PhD, I discovered a lot about myself. One key realization was that I thrive in collaboration. The responsibility I felt toward my students was often what helped me stay focused and carry projects through to the end. I had heard many cautionary tales about writing papers with students, but I was fortunate to work with the most organized and hard-working students anyone could ask for.

First and foremost, I would probably not be defending my PhD without two bachelor's students, Julius Välja and Rasmus Moorits Veski, who chose a group project in the "Natural and Artificial Intelligence" course in 2022. I recalled our first meeting was on 28 October 2022, when we started exploring how to evaluate counterfactual explanations from a human perspective. I am deeply grateful that they remained inspired and motivated for three years and that they chose me as their bachelor's thesis supervisor. Together, we spent countless hours designing our questionnaire (often rethinking it entirely) and yet, thanks to their patience and dedication, we succeeded. We went through the ethics committee process and conducted three rounds of data collection, gathering responses from over 200 participants. Big thanks also to all contributors to this project: especially Kadi Tulver for bringing a valuable psychological and cognitive science perspective and for giving us confidence that we were on the right track; to Eduard Barbu for helping with GPT-4 experiment design and resources; and to Gioacomo Magnifico for contributing to discussions and presenting our poster at AAAI 2025 on short notice due to my visa situation. And, of course, thanks again to Raul Vicente for his guidance and trust. This project truly would not have been possible without all of you.

Another rewarding collaboration that brought me so much joy was supervising Dmytro Shvetsov. He first joined my proposed project through the Neural Networks course on generating counterfactual explanations for images, which later grew into his master's thesis. Around that time, I had the idea that counterfactual explanations could be applied to image segmentation. When I saw his early results, I was reminded of Joonas Ariva's presentation at the Roosta event about the challenges of segmenting kidney tumors. After our first discussion, we immediately saw the potential of working together. Together with Dmytro Fishman and Joonas Ariva, we decided to join forces. The collaboration with the Biomedical Com-

puter Vision Lab turned into a remarkably fast success story, though it required hard work and close coordination from everyone involved. In the end, we managed to put all the pieces together and complete a paper within just a few months, thanks to everyone's tremendous effort and dedication. I also learned a great deal from Dmytro Fishman about how to approach student supervision: with structure, encouragement, and empathy, caring not only about their scientific progress but also about their mental well-being.

A small side note: I am very grateful to Prof. Meelis Kull for organizing the Roosta retreat with the ML, NAIL, NLP, and Biomedical Computer Vision groups. It gave me a much better understanding of what everyone was working on and, perhaps more importantly, helped me overcome my first PhD crisis. At that time, I had been working alone for over a year and felt my motivation fading. During that event, together with Mari-Liis Allikivi, we initiated a discussion about collaboration among PhD students. Verbalizing those challenges and hearing others' perspectives changed my approach entirely. Afterward, I made a list of potential projects and began sharing them with students and colleagues at PhD events, which, as mentioned above, worked really well. Thank you once again, Meelis, for making that happen.

Next, I would like to thank the large consortium of TRUST-AI collaborators: Peter Bosman, Tanja Alderliesten, and Evi Sijben from CWI. It was a pleasure to co-supervise Marije Tromp's thesis and to engage in our many scientific discussions. My gratitude also goes to Nikos Sakkas and the Apintech team for their enthusiasm in tackling the energy consumption prediction problem, and to Gonçalo Reis Figueira and Fábio Silva Moreira from INESC TEC for their coordination and support. Thank you also to our partners at INRIA, LTP, and TAZI for their valuable collaboration throughout the project.

Another thanks to Tarun Khajurija for trusting me to transform his paper into an interpretability paper, for joining me at XAI 2025, and for giving me the best food tour of Istanbul. Working together combined scientific curiosity with friendship.

PhD students and colleagues. I mentioned that we have a lab full of bright, fun people. Therefore, I would like to elaborate on that. I had the best office mates one could wish for: Jesús Javier Reyes Torres, Mary-Ann Kubre, and Ardi Tampuu, for celebrating achievements, bringing snacks to the room, offering comfort in hard times, listening to my news briefings, and showing unwavering support for Ukraine. Your empathy, engagement and willingness to listen helped me process everything that was happening and made me feel less alone. Thanks also to other NAIL colleagues: Tarun Khajurija, Taavi Luik, Taavi Kivisik, Karl Kristjan Kaup, Kristjan Julius Laak, and Madis Vasser. And most importantly, big thanks to Kadi Tulver for always finding time to listen, for wise advice, and for giving valuable feedback on writing in every situation.

Many thanks to the ML group for including us in their events and being such a welcoming lab. Separate thanks to my master's buddies Novin Shahroudi and Viacheslav Komisarenko, my former roommate Joonas Järve, and EXAI collaborator

and apparently puzzle enthusiast Markus Kängsepp.

Thanks also to the "not average but mean" people, Tarun and Mari-Liis, for the best quotes and laughter in the building.

Big thanks to our "girls' chat" for staying strong together in this male-dominated field: Liisa Rätsep, Mari-Liis Allikivi, Mariia Bakhtina, Dzvinka Yarish, Alina Paas, Kadi Tulver, Mary-Ann Kubre, Agnes Luhtaru, Mirjam Paales, Hele-Andra Kuulmets, Maali Tars, and Heili Aavola. A special thanks to Liisa Rätsep for bringing me to Hiiumaa to relax after submitting my thesis.

I am happy and proud of the vibrant PhD community we have built at Delta by organizing regular PhD Events, thanks to the hard work of Mariia Bakhtina, Novin Shahroudi, Taavi Kivisik, Mari-Liis Allikivi, Tarun Khajurija, Vjatseslav Antipenko, Shahla Nobruzova, Ijeoma Faustina Ekeh, and, humbly, myself. Many of our PhD event topics grew out of our personal struggles, and discussing them together made the journey lighter. I still remember the moment I realized that imposter syndrome was not just "my problem", but something almost everyone felt. Several years later, I hear people quoting Kaido Lepik at a PhD graduates panel event. Another memorable event, personally for me, was our panel of supervisors with Raivo, Marlon, and Mark, whose honesty was eye-opening. They reminded us that becoming a professor does not automatically make one an expert in supervision and that learning continues throughout one's career. I especially appreciated Marlon's advice to have honest conversations with supervisors as early as possible. I tried to follow that ever since. Thank you all for your honesty and warmth.

Family. The greatest acknowledgment goes to my husband, Artem. First, thank you for not doing your own PhD at the same time. Otherwise, I am not sure we would have survived! Thank you for your unconditional support, for prioritizing my deadlines and well-being, and for recognizing when I was close to burnout and taking care of everything while I recovered. And to our little hamster, **Lucy**, thank you for bringing light and joy into our home and for gathering such a wonderful community of friends around. Unfortunately, hamsters do not live long enough to see someone through an entire PhD, but they certainly know how to keep you entertained during all those overnight deadlines.

Big thanks to my parents, my brother, and his wife, and all my extended family for keeping each other safe, giving me regular updates, and supporting me. Although given the circumstances, it was somewhat surprising that I turned out to be the one who needed more support and reassurance that everything was as okay as it could be, which only proves how strong you were in shielding me from additional worries.

My Ukrainian Friends. Also big thanks to all Ukrainian friends: Dr. Mariia Bakhtina, Stas Deviatykh, Kateryna Peikova, Alina Paas, Yevheniia Kryvenko, Oleksandra Tkalick, Yuliia Puzanova, Alina Vorontseva, Dr. Vlad Fediukov (congratulations on your PhD), Dr. Olha Kaminska, Vyacheslav Komisarenko (very soon to be a Dr. as well), Dzvinka Yarish, Dr. Kateryna Kubrak, Andrii Tiertyshnyi, Kateryna Porshnieva, Solomiya Branets, Sofiya Demchuk, Olesia Kucheryk,

Dariia Zakharova, Anastansia Babash, Tania Siagailo, and Oleksandr Husiev for your moral support, donations, constant check-ins on each other's families, and for spending time together even if complete silence. A special thanks goes to Dr. Kaido Lepik. Although Estonian, you were one of us in terms of moral support. I have always been grateful for your career advice and friendship. However, I hope that Kateryna Peikova beats you in the table tennis tournaments.

It may seem like a long list of names, but in moments of crisis, it is truly remarkable how people you have known throughout your life come together (organizing transportation, supplies, drones, Ecoflows) and check on one another's families.

Finally, my deepest and most enormous gratitude goes to the **Armed Forces of Ukraine**, who made it possible for my family to stay safe. There were many moments when I doubted whether I should continue my research and PhD, but their strength, bravery, and sacrifice made it possible. I feel sorry for not doing enough. I do not know a single Ukrainian who does not share this feeling. I hope for victory, for Ukraine to remain free and independent, and I sincerely hope that Estonia and the Estonian people will never have to experience a similar struggle.

I am sure I have forgotten to thank someone I intended to, as the past years have been overwhelming in many ways. But I want to say this: during the hardest times, each of you approached me with kindness and care. Because of that, it is now genuinely difficult to imagine pursuing my career anywhere outside Delta. Thank you all for making it so hard to leave.

Слава Україні!

SISUKOKKUVÕTE

Inimkesksete kontrafaktuaalsete seletuste arendamine seletatavas tehisintellektis

Tehisintellekti (TI) kiire areng ja kasvav keerukus on viinud selle laialdase rakendamiseni paljudes olulistes otsustusvaldkondades, nagu tervishoid, rahandus ja haridus. Samas muudab TI mudelite keerukus nende otsustusprotsessid sageli läbipaistmatuks, raskendades kasutajatel ja otsustest mõjutatud isikutel tulemuste mõistmist ja usaldamist. Käesolev doktoritöö tegeleb nende väljakutsetega, arendades ja hinnates kontrafaktuaalseid seletusi, mis aitavad TI otsuseid paremini mõista, tuues välja, mis oleksid minimaalsed vajalikud muudatused, mis võiksid viia teistsuguse tulemuseni. Kontrafaktuaalsed seletused vastavad küsimusele: „Millised minimaalsed sisendi muudatused oleksid muutnud mudeli otsust?“ Selline lähenemine haakub hästi inimliku tunnetetusega, kuna kontrafaktuaalne „mis oleks, kui“ mõtlemine on loomulik viis, kuidas inimesed analüüsivad alternatiivseid teguviise ja nende võimalikke tagajärgi.

Doktoritöös esitatud uurimistöö koosneb neljast omavahel seotud uuringust, mille eesmärk on täiustada kontrafaktuaalseid seletusi, muutes need inimeste kognitiivsete eelistustega sarnasemaks, kergemini hinnatavaks ja praktiliselt rakendatavaks.

Esimeses publikatsioonis tutvustatakse CoDiCE (Coherent Directional Counterfactual Explainer) raamistikku, mille eesmärk on viia kontrafaktuaalsete seletuste otsingu algoritmid inimeste eelistustega paremasse vastavusse. Inimeste eelistused, eelkõige nende teostatavus (feasibility) ja loogiline sidusus (coherence), on raamistikus vormistatud matemaatiliselt läbi kahe mõiste - difuusne kaugus ja suunatud koherentsus. Difuusne kaugus aitab tagada, et kontrafaktuaalsed seletused oleksid teostatavad, tuvastades realistlikud üleminekud läbi andmejaotuse tihedalt seotud piirkondade ning muutes samm-sammulised muudatused praktiliseks ja rakendatavaks. Suunatud koherentsus omakorda tagab, et mudeli poolt soovitatud tunnuste muudatused oleksid intuitiivses vastavuses üldtuntud põhjus-tagajärg seostega. CoDiCE raamistik rakendati Pythonis integreerituna optimeerimistehnikatega, mis võimaldavad andmestruktuurides tõhusalt navigeerida, tagades seejuures koherentsuspiirangute järgmise. Me võrdlesime CoDiCE'i mitmete tuntud meetoditega (DiCE, FACE, Guided Prototypes ja Growing Spheres), kasutades erinevaid andmekogumeid nii klassifikatsiooni (täiskasvanute sissetulek, Saksa krediit, COMPAS, diabeet, rinnavähk) kui ka regressiooni (energia tarbimine) ülesannetes. Katsed näitasid, et CoDiCE genereeris järjepidevalt kontrafaktuaalseid seletusi, mis olid olemasolevate meetoditega võrreldes praktilisemad, realistlikumad ja inimese perspektiivist koherentsemad. Kontrafaktuaalsete seletuste hindamine on valdkonnas veel lahendamata probleem. Selle väljakutsega tegelemiseks lõime mitmekesise andmestiku CounterEval, mis koosneb enam kui 200 inimese hinnangust mitmete seletuslike mõõdikute osas (teostatavus, usaldus, täielikkus,

järjepidevus, keerukus, õiglus, arusaadavus ja üldine rahulolu). Andmestik võimaldas põhjalikult analüüsida inimeste eelistusi ning pakkus empiirilisi tõendeid selle kohta, kuidas erinevad demograafilised rühmad seletusi erinevalt hindavad.

Teises publikatsioonis kasutati CounterEval andmestikku, et uurida, kas suured keelemudelid (LLMid) nagu GPT-4 suudavad matkida inimeste hindamisotsuseid. CounterEval andmetega treenitud LLMid saavutasid kõrge ennustustäpsuse (kuni 85–90%), matkides nii keskmisi inimeste hinnanguid kui ka individuaalseid hindamismustreid. Need tulemused näitavad, et LLMid võivad olla inimhindajate skaleeritavad alternatiivid, lihtsustades oluliselt kontrafaktuaalsete seletuste hindamisprotsessi.

Kolmas publikatsioon uuris, kas üldist rahulolu kontrafaktuaalsete seletustega on võimalik ennustada teiste seletuslike omaduste põhjal. Modelleerides rahulolu hinnanguid skaalal 1–6, saavutasime selgitusmäära $R^2 = 0.75$. Veelgi enam, klassifitseerides rahulolu kolmeks tasemeks (madal, keskmine, kõrge), saavutasime kuni 78% täpsuse. Selgitavate mõõdikute seas olid üldise rahulolu tugevaimad ennustajad teostatavus ja usaldus. Isegi kui need kaks peamist tegurit mudelist välja jätta, suutsid ülejäänud mõõdikud (nt täielikkus, järjepidevus) endiselt seletada 58% varieeruvusest. Lisaks näitas demograafiline analüüs, et meditsiini- või masinõppe taustaga vastajad hindasid seletusi erinevalt, rõhutades seletuste kohandamise vajadust erinevatele kasutajarühmadele.

Lõpuks uuriti **neljandas publikatsioonis** kontrafaktuaalsete seletuste praktilist rakendust meditsiinilise pilditöötluse valdkonnas, arendades selleks COIN (Counterfactual Inpainting) raamistiku. COIN pakub uut lähenemist nõrgalt juhendatud semantilisele segmenteerimisele, mis on eriti kasulik stsenaariumides, kus märgistatud andmeid on raske hankida. Arendasime GAN-il põhineva algoritmi, mis genereerib anomaalseks klassifitseeritud pildile vastava kontrafaktuaalse (normaalse) pildi. Algse ja kontrafaktuaalse pildi erinevusest saab genereerida segmentatsioonimaske. Muutes klassifikaatori väljundi anomaalsest normaalseks ja analüüsides saadud erinevusi, genereerib COIN seletuslikud ja funktsionaalsed segmentatsioonikaardid. COIN meetod andis CAM-põhiste omistusmeetoditega võrreldes paremaid tulemusi ning seda valideeriti sünteetiliste andmete ning Tartu Ülikooli Kliinikumi neerukasvajate andmestikuga. Selline lähenemine võimaldab vähendada radioloogide töökoormust ning aitab siluda klassifikaatoreid, paljastades ekslikest visuaalsetest vihjetest tulenevad probleemid, mida tavapärased meetodid sageli ignoreerivad. Lisaks pakuvad COINi genereeritud kontrafaktuaalsed seletused suuremat ustavust alusmudelile kui traditsioonilised salientsuskaardid, mis võivad kasutajate usaldust eksitada. Kokkuvõttes rõhutavad selle doktoritöö tulemused kontrafaktuaalsete seletuste vastavusse viimise olulisust inimeste kognitiivsete eelistustega. Samuti tutvustab töö erinevaid meetodeid selliste seletuste genereerimiseks, hindamiseks ja rakendamiseks erinevates valdkondades.

7. PUBLICATIONS

CURRICULUM VITAE

Personal data

Full name:	Marharyta Domnich
Date of birth:	01.01.1995
Citizenship:	Ukraine
E-mail:	marharyta.domnich@gmail.com

Education

2021 – 2025	Ph.D. in Computer Science, University of Tartu
2017 – 2020	MSc in Computer Science, University of Tartu (*cum laude*)
2018 – 2019	Exchange studies in Computer Science, Aalto University
2012 – 2016	BSc in Computer Science, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Employment

2021 – 2025	Junior Research Fellow in Computational Neuroscience, University of Tartu
2020 – 2021	Machine Learning Engineer, KappaZeta

Supervision

2024, MSc Thesis	Marije R. Tromp, (sup) Evi Shijben[1]; Marharyta Domnich; Peter A.N. Bosman *"Generating Diverse Counterfactuals through Evolutionary Multi-Objective Optimization"*
2024, MSc Thesis	Dmytro Shvetsov, (sup) Joonas Ariva[1]; Dmytro Fishman; Marharyta Domnich *"Weakly Supervised Segmentation in Medical Imaging: A Counterfactual Approach"*
2024, BSc Thesis	Julius Välja, (sup) Marharyta Domnich[1]; Raul Vicente; Eduard Barbu *"Assessing the Quality of Counterfactual Explanations with Large Language Models"*
2024, BSc Thesis	Rasmus Moorits Veski, (sup) Marharyta Domnich[1]; Raul Vicente; Kadi Tulver *"Measuring Human Preferences in Counterfactual Explanations"*
2022, MSc Thesis	Anton Kostiukhin, (sup) Alexander Kmoch[1]; Marharyta Domnich; Tanel Tamm; Indrek Sünter; Kaupo Voormansik *"Clustering analysis of spatiotemporal Sentinel-2 data of agricultural parcels in Estonia for damaged crop delineation"*
2022, MSc Thesis	Hudson Taylor Lekunze, (sup) Marharyta Domnich[1]; Tambet Matiisen *"NDVI Image Synthesis with Image-to-Image Translation Networks"*

Teaching

Spring 2022, 2023, 2024, 2025	Neural Networks (teaching assistant)
Fall 2022	Natural and Artificial Intelligence Seminar *on Explainability*
Fall 2021	Computational Neuroscience (teaching assistant)

Administrative and professional activities

- Active participation in EU project "Transparent, Reliable and Unbiased Smart Tool for AI" No. 952060 (TRUST-AI) 01.01.2020-31.03.2025
- Bachelor's and Master's theses defense committee member in June 2023, 2024, 2025

[1] Supervisor in charge

ELULOOKIRJELDUS

Isikuandmed

Täisnimi:	Marharyta Domnich
Sünniaeg:	01.01.1995
Kodakondsus:	Ukraina
E-mail:	marharyta.domnich@gmail.com

Haridus

2021 – 2025	Ph.D. arvutiteaduses, Tartu Ülikool
2017 – 2020	MSc arvutiteaduses, Tartu Ülikool
2018 – 2019	Vahetusõpingud arvutiteaduses, Aalto Ülikool
2012 – 2016	BSc arvutiteaduses, Ukraina Riiklik Tehnikaülikool "Igor Sikorsky Kiievi Polütehniline Instituut"

Teenistuskäik

2021 – 2025	Nooremteadur (arvutuslik neuroteadus), Tartu Ülikool
2020 – 2021	Masinõppe insener, KappaZeta

Juhendamine

2024, MSc lõputöö	Marije R. Tromp, Evi Shijben[1]; Marharyta Domnich; Peter A.N. Bosman *"Mitmekesiste kontrafaktuaalide genereerimine evolutsioonilise mitme-eesmärgilise optimeerimise abil"*
2024, MSc lõputöö	Dmytro Shvetsov, Joonas Ariva[1]; Dmytro Fishman; Marharyta Domnich *"Nõrgalt juhendatud segmenteerimine meditsiinilises pilditöötluses: kontrafaktuaalne lähenemine"*
2024, BSc lõputöö	Julius Välja, Marharyta Domnich[1]; Raul Vicente; Eduard Barbu *"Kontrafaktuaalsete seletuste kvaliteedi hindamine suurte keelemudelite abil"*
2024, BSc lõputöö	Rasmus Moorits Veski, Marharyta Domnich[1]; Raul Vicente; Kadi Tulver *"Inimeste eelistuste mõõtmine kontrafaktuaalsete seletuste puhul"*
2022, MSc lõputöö	Anton Kostiukhin, Alexander Kmoch[1]; Marharyta Domnich; Tanel Tamm; Indrek Sünter; Kaupo Voormansik *"Spatiotemporaalsete Sentinel-2 andmete klasteranalüüs Eesti põllumaadel kahjustatud põllukultuuride määramiseks"*
2022, MSc lõputöö	Hudson Taylor Lekunze, Marharyta Domnich[1]; Tambet Matiisen *"NDVI piltide süntees kujutisest-kujutiseks teisendusvõrkude abil"*

[1] Vastutav juhendaja

Õppetöö

Kevad 2022, 2023, 2024, 2025	Tehisnärvivõrgud (õppeassistent)
Sügis 2022	Tehisliku ja loomuliku mõistuse seminar seletavuse teemal
Sügis 2021	Arvutuslik neuroteadus (õppeassistent)

Administratiivsed ja professionaalsed tegevused

- Aktiivne osalemine EL projektis "Transparent, Reliable and Unbiased Smart Tool for AI"nr 952060 (TRUST-AI), 01.01.2020–31.03.2025
- Bakalaureuse- ja magistritööde kaitsmiskomisjoni liige juunis 2023, 2024, 2025

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.
59. **Maksym Del.** Multilingual and Multi-Domain Representational Patterns Across Trpansformer-Based Models. Tartu 2024, 131 p.
60. **Kristo Raun.** Adaptive Out-of-order Handling in Streaming Conformance Checking. Tartu 2024, 118 p.
61. **Toivo Vajakas.** Towards integration of mobile network data into analyzing human mobility. Tartu 2024, 103 p.
62. **Katsiaryna Lashkevich.** Data-Driven Analysis and Optimization of Waiting Times in Business Processes. Tartu 2024, 169 p.
63. **Alejandra Duque-Torres.** Classifying, Constraining and Ranking Metamorphic Relations. Tartu 2025, 159 p.

64. **Mariia Bakhtina.** A Method for Information Security and Privacy Management in Smart Solutions. Tartu 2025, 199 p.
65. **Andre Tättar.** Multilingual Machine Translation for Under-Resourced Languages. Tartu 2025, 170 p.
66. **Mahmoud Shoush.** Prescriptive Process Monitoring Under Uncertainty and Resource Constraints. Tartu 2025, 178 p.
67. **Alireza Akhavi Zadegan.** A Multimodal approach for refining Mapping and Localization by Integrating Generative AI and Pedestrian-Centric Data. Tartu 2025, 147 p.
68. **Eerik Muuli.** Automating the assessment and feedback processes in IT teaching – improving creation and maintenance from the teaching staff perspective. Tartu 2025, 196 p.
69. **Kateryna Kubrak.** Towards User-Centered Prescriptive Process Monitoring Systems. Tartu 2025, 151 p.
70. **Zhigang Yin.** Computing and Sensing in a Smart Ring. Tartu 2025, 251 p.
71. **Abdul-Rasheed Olatunji Ottun.** Practical Trustworthy Artificial Intelligence with Human Oversight. Tartu 2025, 239 p.
72. **Sander Mikelsaar.** Analysis and Optimization of Iteratively Decodable Codes. Tartu 2025, 146 p.