Counterfactual Explanations and Algorithmic Recourse for Trustworthy AI

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Wednesday, 25 February 2026 at 17:30

by

Patrick ALTMEYER

MA Hons Economics, The University of Edinburgh, United Kingdom
Master Degree in Economics and Finance, Barcelona School of Economics, Spain
Master Degree in Data Science, Barcelona School of Economics, Spain

born in Düsseldorf, Germany

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Dr. ir. C.C.S. Liem	Delft University of Technology, *promotor*
Prof. dr. A. van Deursen	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. ir. J.H. Kwakkel	Delft University of Technology
Prof. dr. I.P.P. van Lelyveld	Vrije Universiteit Amsterdam, De Nederlandsche Bank
Dr. M. Mitchell	Hugging Face, Inc. (United States)
Prof. dr. M. Pechenizkiy	Eindhoven University of Technology
Prof. dr. F.A. Oliehoek	Delft University of Technology, reserve member
Other members:	
Dr. F. Barsotti	Delft University of Technology, ING Bank N.V.

The work presented in this thesis has been conducted as part of the AI for Fintech Research Lab at ING, operating under the Innovation Center for Artificial Intelligence (ICAI) flag.

Keywords: Artificial Intelligence, Trustworthy AI, Counterfactual Explanations, Algorithmic Recourse

Printed by: ProefschriftMaken

Cover by: Patrick Altmeyer

The author set this thesis using Quarto: https://github.com/quarto-tudelft.

ISBN 978-94-6518-234-6

An electronic copy of this dissertation is available at https://repository.tudelft.nl/.

To Dani, my parents, Yannick, Alena and my whole family.

Algorithms do not listen, nor do they bend.

Cathy O'Neil

TABLE OF CONTENTS

SUMMARY

Many of the most celebrated recent advances in artificial intelligence (AI) have been built on the back of highly complex and opaque models that need little human oversight to achieve strong predictive performance. But while their capacity to recognize patterns from raw data is impressive, their decision-making process is neither robust nor well understood. This has so far inhibited trust and widespread adoption of these technologies. This thesis contributes to research efforts aimed at tackling these challenges, through interdisciplinary insights and methodological contributions.

The principle goal of this work is to contribute methods that help us in making opaque AI models more trustworthy. Specifically, we aim to (1) explore and challenge existing technologies and paradigms in the field; (2) improve our ability to hold opaque models accountable through thorough scrutiny; and, (3) leverage the results of such scrutiny during training to improve the trustworthiness of models.

Methodologically, the thesis focuses on counterfactual explanations and algorithmic recourse for individuals subjected to opaque AI systems. We explore what type of real-world dynamics can be expected to play out when recourse is provided and implemented in practice. Based on our finding that individual cost minimization–a core objective in recourse–neglects hidden external costs of recourse itself, we revisit yet another established objective: namely, that explanations should be plausible first and foremost. Our work demonstrates that a narrow focus on this objective can mislead us into trusting fundamentally untrustworthy systems. To avoid this scenario, we propose a novel method that aids us in disclosing explanations that are maximally faithful, that is consistent with the behavior of models. This not only allows us to assess the trustworthiness of models, but also improve it: we show that faithful explanations can be used during training to ensure that models learn plausible explanations.

Finally, we also critically assess efforts towards trustworthy AI in the context of modern large language models (LLM). Specifically, we cast doubt on recent findings and practices presented in the field of mechanistic interpretability and caution our fellow researchers in this space against misinterpreting and inflating their findings.

In summary, this thesis makes cutting-edge research contributions that improve our ability to make opaque AI models more trustworthy. Beyond our core research contributions, this thesis makes substantial contributions to open-source software. Through various software packages that we have developed, we make our research and that of others more accessible.

SAMENVATTING

Veel van de meest geprezen recente ontwikkelingen op het gebied van kunstmatige intelligentie (AI) zijn gebouwd op basis van zeer complexe en intransparante modellen die weinig menselijk toezicht nodig hebben om sterke voorspellende prestaties te behalen. Maar hoewel hun vermogen om patronen uit ruwe data te herkennen indrukwekkend is, is hun besluitvormingsproces noch robuust noch goed begrepen. Dit heeft tot nu toe het vertrouwen in en de wijdverspreide adoptie van deze technologieën belemmerd. Dit proefschrift draagt bij aan onderzoeksinspanningen die gericht zijn op het aanpakken van deze uitdagingen, door middel van interdisciplinaire inzichten en methodologische bijdragen.

Het hoofddoel van dit werk is om methoden bij te dragen die ons helpen met het betrouwbaarder maken van intransparante AI-modellen. Specifiek streven we ernaar om (1) bestaande technologieën en paradigma's in het veld te verkennen en te bevragen; (2) ons vermogen te verbeteren om intransparante modellen verantwoordelijk te houden door middel van grondige inspectie; en, (3) de resultaten van dergelijke inspectie tijdens de modeltraining te benutten om de betrouwbaarheid van modellen te verbeteren.

Methodologisch richt het proefschrift zich op 'counterfactual explanations'–contrafeitelijke verklaringen–en 'algorithmic recourse'–algoritmische hulpmiddelen–voor individuen die worden blootgesteld aan intransparante AI-systemen. We onderzoeken welke dynamieken in de praktijk kunnen worden verwacht wanneer algorithmic recourse worden aangeboden en geïmplementeerd. Gebaseerd op onze bevinding dat individuele kostenminimalisatie—een kerndoelstelling bij recourse—verborgen externe kosten van recourse zelf negeert, heroverwegen we nog een ander algemeen aanvaard doel: namelijk dat de uitleg van algoritmische beslissingen in de eerste plaats plausibel moeten zijn. Ons werk toont aan dat een dergelijke interpretatie van uitlegbaarheid ons kan misleiden om fundamenteel onbetrouwbare systemen te vertrouwen.

Om dit scenario te voorkomen, stellen we een nieuwe methode voor die ons helpt bij het vinden van verklaringen die zo goed mogelijk aansluiten bij het daadwerkelijke gedrag van modellen. Dit stelt ons in staat de betrouwbaarheid van AI niet alleen te beoordelen, maar ook te verbeteren: we laten zien dat waarheidsgetrouwe verklaringen tijdens de training kunnen worden gebruikt om te verzekeren dat modellen plausibele verklaringen leren.

Tot slot kijken we kritisch naar het vraagstuk van betrouwbare AI in de context van moderne grote taalmodellen (LLMs). Specifieker stellen we vragen bij recente

bevindingen en praktijken in het veld van mechanistische interpreteerbaarheid, en waarschuwen we onze collega-onderzoekers in dit gebied voor het verkeerd interpreteren en opblazen van hun resultaten.

Samenvattend levert dit proefschrift baanbrekende onderzoeksbijdragen die ons vermogen verbeteren om intransparante AI-modellen betrouwbaarder te maken. Naast onze kernonderzoeksbijdragen levert dit proefschrift substantiële bijdragen aan open-source software. Door middel van verschillende softwarepakketten die we hebben ontwikkeld, maken we ons onderzoek en dat van anderen toegankelijker.

1

INTRODUCTION

Recent developments in artificial intelligence (AI) have largely centered around **representation learning**: instead of relying on features and rules that are carefully hand-crafted by humans, modern machine learning (ML) models are tasked with learning representations directly from data to make predictions (Goodfellow, Bengio, and Courville 2016)—this typically involves optimizing these representation to achieve narrow training objectives like predictive accuracy. Modern advances in computing have made it possible to provide such models with ever-growing degrees of freedom to achieve that task, which frequently allows them to outperform traditionally more parsimonious models. While this branch of AI has certainly not been the only active field of research, it is arguably the one that has attracted the highest levels of public attention and investment over the past decade. This trend has been fuelled by increasingly bold promises that "big data leads to better [...] decisions" (McAfee et al. 2012), companies embracing "machine learning will be the big winners of tomorrow" (Tank 2017) and that, ultimately, "[AI] could massively accelerate scientific discovery and innovation well beyond what we are capable of doing on our own" (Altman 2025).

Unfortunately, the models underlying all these developments learn increasingly complex and highly sensitive representations that humans can no longer easily interpret. This trend towards complexity for the sake of performance has come under serious scrutiny in recent years. One important challenge arising from high sensitivity is model robustness: at the very cusp of the deep learning (DL) revolution, Szegedy et al. (2014) showed that artificial neural networks (ANN) are sensitive to adversarial examples (AEs): perturbed versions of data instances that yield vastly different model predictions despite being "imperceptible" in that they are semantically indifferent from their factual counterparts. Even though some partially effective mitigation strategies have been proposed—most notably **adversarial training** (Goodfellow, Shlens, and Szegedy 2015)—truly robust deep learning remains unattainable even for models that are considered "shallow" by today's standards (Kolter 2023).

Another obvious challenge of increased complexity is our own lack of human understanding with respect to the decision logic underlying these models: as one recent work puts it, "nobody understands deep learning" (Prince 2023). This, too, has attracted much criticism: O'Neil (2016) pointed to the dangers of deploying such opaque models in the real world; Buolamwini and Gebru (2018) uncovered hidden biases of supposedly 'neutral algorithms' and Rudin (2019) argued against using opaque models altogether. On the other end of the spectrum, the "black-box" challenge (as it is sometimes called) has attracted an abundance of research on **explainable AI** (XAI), a paradigm that focuses on the development of tools to derive (post-hoc) explanations from complex model representations. Such explanations should mitigate a scenario in which practitioners deploy opaque models and blindly rely on their predictions. Effective XAI tools hold the promise of not only aiding us in monitoring models, but also providing recourse to individuals subjected to them.

Part of the problem is that the high degrees of freedom provide room for many solutions that are locally optimal with respect to narrow objectives (Wilson 2020).[1] Indeed, recent work on the so-called "lottery ticket hypothesis" suggests that modern neural networks can be pruned by up to 90% while preserving their predictive performance (Frankle and Carbin 2019). Similarly, Zhang et al. (2021) showed that state-of-the-art neural networks are expressive enough to fit randomly labeled data. Thus, looking at the predictive performance alone, the solutions may seem to provide compelling explanations for the data, when in fact they are based on purely associative, semantically meaningless patterns.

While we believe that for large enough models, bullet-proof explainability remains as unattainable as robustness, the contributions of this thesis demonstrate that XAI tools can help us to not only shed light on the solutions space, but tame it. We will show that is important to not simply seek and isolate model explanations that satisfy us, but rather think of explanations as distributional quantities that depend on both the underlying data and the model. By faithfully presenting the whole spectrum of these distributions and inducing models to be aligned with the subset of explanations that humans consider meaningful, XAI can make fundamental progress towards trustworthy AI.

1.1. TRUSTWORTHY ARTIFICIAL INTELLIGENCE

Trustworthy AI is a relatively novel term spanning a broad field of research. It covers a range of subtopics including fairness, ethics, societal impact and explainability. Varshney (2022) represents the first concerted effort towards unifying and defining related concepts in a single self-contained resource. The urgency for this kind of effort and the field as whole first crystallized in the early 2010s when both industry and regulators began using AI to process the vast amounts of data afforded by the

[1] We follow the standard ML convention, where "degrees of freedom" refer to the number of parameters estimated from data.

digitalization of society. From recommender systems used by tech giants to tailor consumer advertisements to natural language processing (NLP) used by central banks to monitor economic sentiment—everyone has been eager to innovate in this space. But aside from innovation and progress, novel and disruptive technologies also generally present society with new challenges.

O'Neil (2016) was among the first to point out some of these challenges in her influential book 'Weapons of Maths Destruction'. Backed by numerous real-world examples, O'Neil (2016) makes a striking case for why should be very careful and even skeptical of using opaque algorithms to organize society. At times when some governments have chosen to co-operate with the tech industry to monitor and organize sensitive social security data of their constituents and even deploy AI in the context of nuclear security (OpenAI 2025; Field 2025), O'Neil's warnings seem more relevant than ever. AI is not inherently good or bad, but it also will not hold itself accountable for any real-world consequences—good or bad. That remains our responsibility and ongoing efforts towards trustworthy AI play an important role in fulfilling it.

While that responsibility has mostly been shunned by those intent on moving fast and breaking things[2], we remain cautiously optimistic about improving things from the inside, much in line with Varshney (2022). Conscious of the "increasingly sociotechnical nature of machine learning", he defines trustworthy AI in terms of its interaction with humans and society at large. Since we will refer back to this at times, we restate his definition of trustworthy AI here:

Definition 1.1 (Trustworthy AI)**.** For an AI system to be considered *trustworthy*, it needs to fulfill the following criteria:

1. Achieve *basic performance* at the task its intended to be used for.
2. Achieve this performance *reliably*, i.e. safely, fairly and robustly.
3. Facilitate *human interaction* through predictability, understandability and ideally transparency.
4. Be *aligned* with our agenda.

Even though modern AI systems generally fail to comply with this definition and "corporations do not trust artificial intelligence and machine learning in critical enterprise workflows" (Varshney 2022), Definition 1.1 provides goal posts that are not out of reach. In fact, we would argue that at least in those cases where we are sufficiently tempted to use AI today, *basic performance* is usually not an issue. In this work, we typically work on one or more of the remaining three criteria under the assumption that some complex AI tool is preferable over a more simple solution in terms of performance. We are well aware that this assumption does not always hold, and simple tools are often preferable (Rudin 2019). Still, we believe that complex,

[2] Until 2014, "move fast and break things" was part of Meta's official motto (then still operating under the name of Facebook). The phrase has been used to characterize the broader tech industry (see, for example, Vardi (2018)).

opaque models are here to stay and hence we aim to contribute towards making them more trustworthy.

Before we zoom in on the subdomain of trustworthy AI that is most relevant to this thesis, it is worth pointing out that contributions in this space, especially regarding sociotechnical problems, have come from a variety of disciplines and communities (Buszydlik et al. 2025). These communities are not generally focused on providing technical solutions for AI, which stands in contrast to the domain of Explainable AI introduced in the following section (Buszydlik et al. 2024).

1.2. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Considering Definition 1.1, our work contributes primarily to improving AI's potential for criterium 3: *human interaction*. Specifically, most of our methodological contributions are geared towards fostering predictability and understandability of models through the means of Explainable AI (XAI). The field of XAI is concerned with creating methods that improve the explainability of models and thus foster human interaction and trust (Arrieta et al. 2020). This subfield of trustworthy AI is active and large. We will once again refrain from attempting to provide a detailed general introduction to the topic and instead refer readers to Molnar (2022) for a comprehensive overview. For the remainder of this work, it suffices to understand that our methodological contributions largely fall into the category of *post-hoc*, *local*, explanations for opaque, supervised models. This family of models most notably includes ANNs, but also other popular ML models including random forests, XGBoost and Support Vector Machines (SVM). We distinguish opaque *explainable* models from inherently *interpretable* models (Rudin 2019). The latter category includes models that are interpretable by design, such as linear regression, logistic regression and shallow decision trees (Molnar 2022).

Local explanations are local in that they apply to individual samples, sometimes synonymously referred to as instances or inputs. Specifically, they explain the mapping from individual inputs to predictions for opaque models (Molnar 2022). Among the most popular local explanation methods are LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017) and counterfactual explanations (CE) (Wachter, Mittelstadt, and Russell 2017). LIME and SHAP are closely related in that they both use locally additive, linear and interpretable surrogate models to explain the predictions made by the opaque model. SHAP, in particular, has gained huge popularity among researchers and practitioners, likely due to being solidly rooted in game theory and ready availability of multiple open-source software implementations (Molnar 2022). Both LIME and SHAP rely on input perturbations in the local neighborhoods of individual instance to construct the surrogate explanation model, which makes them vulnerable to adversaries (Slack et al. 2020). The reliance on surrogate models is one key feature that distinguishes LIME and SHAP from counterfactual explanations, the method we focus on in this work.

1.3. COUNTERFACTUAL EXPLANATIONS

Instead of locally approximating the behavior of opaque models, **counterfactual explanations** work under the premise of evaluating perturbed inputs directly with respect to the opaque model. Specifically, valid counterfactuals are perturbed inputs that yield some pre-determined change in the prediction of the model. This makes the interpretation of counterfactual explanations intuitive and straightforward: perturbations to individual inputs tell us directly what type of feature changes would have been necessary to yield some desired prediction (Molnar 2022).

Typically, counterfactuals are generated with the objective to minimize those necessary changes, which is how the counterfactual search objective was originally framed in the seminal work by Wachter, Mittelstadt, and Russell (2017). This makes sense, if we think of CE as a means to provide algorithmic recourse (AR) to individuals adversely affected by automated decision-making (ADM) systems (a.k.a. "Weapons of Math Destruction" (O'Neil 2016)). In this context, minimal changes to features can be thought of as minimal costs to individuals who need to implement recourse to change negative into positive predictions and outcomes. But as we will demonstrate in this thesis, minimizing costs to individuals in this way neglects the downstream effects that individual recourse can be expected to have on the broader group stakeholders. Still, proximity—in terms of minimal distance or cost—is one of the core desiderata for counterfactuals (Verma et al. 2022; Karimi et al. 2021).

Of course, even minimal feature changes may be infeasible in practice: individuals cannot change their height, age or ethnicity, for example, but if a model is sensitive to these features, then unconstrained counterfactuals will inevitably reflect this and the resulting recourse recommendations will not be actionable. **Actionability** is therefore another key desideratum for CE and AR that has received attention in the literature (Ustun, Spangher, and Liu 2019). We will see that at least for counterfactual explainers that rely on gradient-based optimization, it is straight-forward to respect actionability constraints. But while this is fortunate news with respect to actionable recourse, we will also argue that actionability constraints should really be addressed before the inference stage, during model training. For models with high degrees of freedom, this is of course not trivial.

By now it may already be obvious that counterfactual explanations are not unique. After all, we can perturb features in many, possibly infinite ways to achieve some desired prediction. Suppose, for example, that we have an opaque model that predicts whether individuals qualify for a loan to purchase a home. In this case, the outcome of interest is binary from the perspective of individuals affected by the model: $\{0 := \text{empty hands}, 1 := \text{homeowners}\}$. Assuming the model is any conventional classification model, there exist infinitely many unique counterfactual states on either side of the decision boundary. This inherent multiplicity of explanations has been described as a limitation of CE in some places (Molnar 2022), presumably because it challenges us to form a feasible and desirable subset of explanations. Much of the existing work in this field has indeed been focused on designing methodologies for generating counterfactuals that meet certain desiderata. Some have

explicitly embraced multiplicity of explanations and argued that its desirable to end up with a diverse set of counterfactuals (Mothilal, Sharma, and Tan 2020). In the context of algorithmic recourse, this corresponds to offering individuals a menu of recourse recommendations to choose from according to their own preferences.

Apart from proximity and diversity, various works have proposed methods aimed at ensuring **plausibility** of explanations (Joshi et al. 2019; Poyiadzi et al. 2020; Schut et al. 2021). The guiding principle is to generate counterfactuals that are close to the data manifold in the target domain. Since the target domain is generally different from the factual domain—that is the domain the instance originally belongs to—any improvements with respect to plausibility inevitably decrease proximity: a counterfactual cannot be close to its factual and the target manifold at the same time. These types of trade-offs between different desiderata are not uncommon, although fortunately different desiderata also tend to complement each other. Plausibility, for example, has also been linked to robustness of counterfactuals (Artelt et al. 2021), where explanations are considered as robust to the extent that they remain valid if the model or data changes (Pawelczyk et al. 2023). Robustness of counterfactuals has in turn been linked to diversity (Leofante and Potyka 2024).

Navigating the sheer amount of desiderata for CE and their interplay can be challenging: depending on the context, domain and even individual users, one may need to optimize for one desideratum at the cost of another. In this thesis, we offer one guiding principle that should help researchers and practitioners in this respect. Specifically, we argue and demonstrate that counterfactual explanations should first and foremost be faithful to the model in question. In other words, counterfactuals should be consistent with what the model has learned about the underlying problem and data. **Faithfulness** has previously been largely ignored by researchers, but we demonstrate that neglecting this desideratum can lead to undesirable outcomes. It is, for example, generally possible to generate plausible counterfactuals for even the most fragile and untrustworthy models that were optimized solely for accuracy. But if these counterfactuals do not faithfully explain model behavior, they are not only useful but potentially misleading, instilling a false sense of trust in poorly trained models.

1.4. TRUSTWORTHY AI IN THE AGE OF LLMS

Existing challenges with respect to the trustworthiness of opaque AI models have become more pressing in recent years as the scale and potential impact of AI systems on society has increased in the age of LLMs. Following the release of ChatGPT, even some of the most influential and respected AI researchers were in such awe that they publically expressed concern around our capability to control these systems, spurring an active debate and research on AI safety and explainability (Future of Life Institute 2023b). An emerging line of research in this context is mechanistic interpretability, which aims to shed light on the inner workings of vast neural networks.

There have been promising advances in this field that aid us in understanding, monitoring and controlling the tools we have so readily deployed on society (Bereska and Gavves 2024). Unfortunately though, there has also been a tendency in some circles to jump from interpretability findings to premature conclusions about AGI. As a final research contribution of this thesis, we critically assess this trend and call for greater caution and modesty in interpreting and presenting such findings.

1.5. GOALS AND RESEARCH QUESTIONS

As stated earlier, the principal goal of this work is to contribute methods that help us in making opaque AI models more trustworthy. Since the field of trustworthy AI is still relatively young, it is important that research and any related software are made widely and openly accessible to other researchers and practitioners.

1.5.1. GOALS

The principal goals of this thesis are as follows:

1. Explore and challenge existing technologies and paradigms in trustworthy AI, in particular with respect to explainability.
2. Improve our ability to hold complex machine learning models accountable through novel methods that facilitate thorough scrutiny.
3. Leverage the results of such scrutiny to aid us in building models that are inherently more trustworthy.

General principles that have played a role in achieving all of these goals include a strong adherence to best practices for producing reproducible and accessible research, as well as open-source software. The remainder of Section 1.5 dives deeper into more granular research questions that have grown out of these principal goals.

1.5.2. COUNTERFACTUAL EXPLANATIONS AND OPEN-SOURCE

Open-source software implementations of LIME and SHAP have contributed to the popularity of these methods (Molnar 2022) and we strive to achieve the same outcome for counterfactual explanations. Specifically, we aim to make existing work in the field readily available and in doing so, we hope to inform our own research about any existing gaps, challenges or open questions. Ultimately, it is our goal to contribute methodological advances accompanied by state-of-the-art open-source software that enable researchers and practitioners to not only better understand the behavior of opaque AI models, but also use that understanding in order to improve their trustworthiness.

To achieve this goal, we begin our research trajectory with the following question:

Thesis Research Question 1.1: Counterfactual Explanations and Open-Source

What are counterfactual explanations, why are they useful for trustworthy AI and what gaps are there in the existing open-source software landscape?

1.5.3. DYNAMICS OF CE AND AR

As part of answering this question, in Chapter 2 we introduce a novel comprehensive, extensive and highly performant software implementation for generating counterfactual explanations in the Julia programming language (Bezanson et al. 2017; Altmeyer, Deursen, and Liem 2023a). This is a first important step towards facilitating human interaction with opaque AI in the context of this thesis (Definition 1.1). The fast performance of Julia and our package allows us to explore previously untapped challenges that relate to the dynamics of counterfactual explanations (Verma et al. 2022). In particular, we ask ourselves:

Thesis Research Question 1.2: Dynamics of CE and AR

What dynamics are generated when off-the-shelf solutions to CE and AR are implemented in practice?

1.5.4. PLAUSIBILITY AND FAITHFULNESS

In consideration of Definition 1.1, we specifically wonder if by facilitating human interaction we risk creating adverse effects on other aspects of trustworthiness including basic model performance and reliability. Answering these questions requires computationally expensive simulations that involve repeatedly generating CE and AR and (re-)training machine learning models. Findings from such simulations help us to uncover consequences that were difficult to predict when designing initial objectives for individual recourse. Our work on this question makes it clear that a narrow focus on minimizing costs to individuals can create dynamics that are costly to other individuals and stakeholders (Chapter 3). To avoid such endogenous dynamics, CE and AR need to be consistent with the data-generating process, which we have referred to above as 'plausible'. Since existing work on generating plausible counterfactuals typically involve surrogate models that are not strictly needed to generate valid CE, we wonder:

Thesis Research Question 1.3: Plausibility and Faithfulness

Can we generate plausible counterfactuals relying only on the opaque model itself?

1.5.5. COUNTERFACTUAL TRAINING

We find that this is not only possible, but also constitutes a cleaner and more principled approach towards explaining models through counterfactuals. It mitigates the risk of entangling the behavior of the opaque model with the surrogate. We demonstrate that only faithful explanations enable us to distinguish trustworthy from untrustworthy models (Chapter 4). We consider this as one of the key steps towards truly understanding the behavior of opaque models and thus fostering meaningful human interaction (Definition 1.1). It allows us to ask the following question:

Thesis Research Question 1.4: Counterfactual Training

How can we leverage faithful counterfactual explanations during training to build more trustworthy models?

Suppose we have trained some opaque model that achieves good basic (predictive) performance, but faithful explanations reveal that it is untrustworthy. In other words, the supervised model excels at its narrow discriminative objective by making predictions based on associations in the data that are not meaningful to humans. Knowing that this model is not trustworthy is useful in and of itself, but in lack of a more principled framework to act on this information it creates a dilemma: should we still go ahead and use the model or discard it in favor of a more trustworthy, but possibly less performant alternative. Ideally, we would like to have the best of both worlds by improving the trustworthiness of the performant model. Since we typically have a pre-defined notion of meaningful explanations for data, we wonder if it is possible to use faithful explanations as feedback for models during training. Our work on this question directly targets the alignment aspect of Definition 1.1 and indirectly improves all other aspects (Chapter 5).

1.5.6. TRUSTWORTHY AI AND LLMS

Even though our work remains focused on contributions to core research questions in the field of CE and AR, we are not oblivious to the advancements and potential societal impacts of LLMs. It is therefore natural to ask ourselves to what extent existing work on trustworthy AI (including our own) can play a role in better understanding the behavior these models. In particular, we ask:

Thesis Research Question 1.5: Trustworthy AI and LLMs

Can we explain the predictions of LLMs and do recent findings from mechanistic interpretability really hint at AGI?

The first part of this question is naturally aligned with the broader scope of this work. The second part is a reaction to concerning trends and tendencies of some fellow researchers to make unscientific claims about AGI based on questionable

evidence. We find it necessary to distance ourselves from such practices and to caution other researchers against it, because we believe they dampen the credibility of otherwise valid and valiant efforts towards improved trustworthiness through mechanistic interpretability (Chapter 6).

1.6. RESEARCH METHODOLOGY

This work has been predominated by quantitative methods and software development. Development has often informed research and vice-versa.

1.6.1. QUANTITATIVE METHODS

All chapters contain descriptions and mathematical expositions of specific quantitative methods, as well as computational experiments involving both synthetic, vision and real-world tabular datasets. Since counterfactual explanations involve a counterfactual search objective, optimization—in particular stochastic gradient-based optimization—has been the main quantitative method that unites Chapter 2 to Chapter 5. Across these chapters we also make use of simulations (Chapter 3), bootstrapping (most notably Chapter 4 and Chapter 5), statistical divergence measures, confidence intervals and hypothesis testing. We also borrow and adapt established methods from contrastive learning (Chapter 4 and Chapter 5), robust (adversarial) learning (Chapter 5), and conformal prediction (Chapter 4). Chapter 5 also involves a formal mathematical proof. In Chapter 6, we employ tools from mechanistic interpretability for LLMs such as linear probes and propose a specific hypothesis test. All of our research works involve deep learning and other machine learning models. Quantitative methods that have not or only indirectly been employed in any of the chapters but nonetheless played an important role in our research and development process include: Laplace approximation, Bayesian deep learning, (variational) autoencoders, decision trees and tree-based algorithms. Finally, we have made heavy use of multiprocessing and multithreading to run extensive computational experiments as part of Chapter 4 and Chapter 5.

1.6.2. INTERDISCIPLINARY RESEARCH

During his previous employment as an economist at the Bank of England (Appendix C), the author of this dissertation realized that despite a growing appetite for AI, monetary policymakers were rightly skeptical of models they cannot fully understand nor trust—after all, the decisions made by central banks affect the lives of entire populations. This background has helped shape much of the work in this thesis, because it has enabled the author to consider certain problems from a unique interdisciplinary angle. Some chapters of this thesis are indeed interdisciplinary in that they are characterized by a bridging of financial and economical expertise and machine learning expertise: Chapter 3 essentially reformulates algorithmic recourse

as a scarce resource over which multiple stakeholders compete; Chapter 6 involves elements and data from economics, finance and psychology, driven by diverse academic and professional backgrounds of the group of authors; and, elements of this thesis including Chapter 2, Chapter 4 and Chapter 6 were presented during invited talks at central banks and other financial institutions such as the Bank of England, De Nederlandsche Bank and the Verbond van Verzekeraars.

A specific example that plays a role in the context of Chapter 6 should help to illustrate how this work has benefited from interdisciplinary perspectives: the concept of "emergence" in complex AI systems, which has been tied to AGI in some places. One can draw a parallel to the "emergence" of asset price bubbles in financial markets (which are complex systems): asset bubbles involve prolonged and often dramatic increases in prices, far beyond the fundamental value of assets. While they may involve rational and predictable behavior of individual economic agents (Brunnermeier 2016), their emergence is notoriously hard to explain, and they typically create substantial economic damage (Mishkin et al. 2008). Economists have proposed no shortage of models and methods to explain and detect bubbles, but to the best of our knowledge none has ever attributed such asset price dynamics to some latent intelligence of markets.

1.6.3. FAIR DATA AND SOFTWARE MANAGEMENT

Throughout this project, we have made an effort to comply with FAIR data principles (Wilkinson et al. 2016). All of our research papers and the accompanying code bases are maintained in version-controlled repositories, which are organized and documented according to best practices either as a Julia project or—in most cases—a fully-fledged package (see Table 1.1 below). In both cases, Julia's package manager `Pkg.jl` handles all dependencies as specified in the `Project.toml` files contained in the repositories. Projects can be forked and cloned to local machines, while packages can be installed from running Julia sessions using `Pkg.jl`. We use Zenodo and 4TU.ResearchData to permanently archive research results on the web and create digital object identifiers (DOI) for individual releases of the various code bases. These releases are generally managed using semantic version control (SVC). Relevant DOIs specific to the individual papers are listed in Table 1.1. All of our experiments rely on publically available datasets, so in terms of new data, besides the software itself we only release our research results. Consistent with TU Delft's Open Access policy, all research papers included in this thesis have been made freely available on the pure.tudelft.nl repository.

1.7. OUTLINE AND CONTRIBUTIONS

So far we have presented the overarching topics and questions that have shaped this work with occasional references to where they appear in the remainder of this thesis. In this final section of the introduction, we provide an outline of what

follows along with detailed descriptions of our contributions. The body of this thesis consists of independent and original research papers that have been peer-reviewed and published (Chapter 2 to Chapter 6). They each individually address different thesis research questions outlined above and contribute to varying aspects of Definition 1.1. Unless explicitly stated otherwise, the papers are included in their original form to ensure their integrity. Only minor modifications have been made if any at all.

In Chapter 2, we present `CounterfactualExplanations.jl`: a package for generating Counterfactual Explanations (CE) and Algorithmic Recourse (AR) for opaque machine learning models in Julia. We discuss the usefulness of CE for explainable AI and demonstrate the functionality of the package. The package is straightforward to use and designed with a focus on customization, extensibility and performance. It is the de facto go-to place for counterfactual explanations and among the most prominent packages for XAI in Julia: at the time of writing, the package has received well over 100 stars on GitHub—somewhat higher but broadly in the same range as ExplainableAI.jl and ShapML; the package also counts over ten contributors, was the main target of a successful Julia Seasons of Contributions project and has been presented to the developer community in main talks at JuliaCon 2022 and 2024.

> We have developed extensive research software in Julia (Bezanson et al. 2017), utilizing other languages including R, Python and Lua in supporting functions. A result of this—and a major contribution of this thesis—is the Taija package ecosystem for trustworthy AI in Julia (67 followers and 24 contributors on GitHub). It includes packages for model explainability (CounterfactualExplanations.jl, predictive uncertainty quantification (ConformalPrediction.jl [142 stars], LaplaceRedux.jl [47 stars]), Bayesian deep learning (LaplaceRedux.jl) and energy-based models (JointEnergyModels.jl). Additionally, there are number of meta packages that ship supporting functionality for the core packages: visualizations (TaijaPlotting.jl), datasets for testing and benchmarking (TaijaData.jl) and parallelization (TaijaParallel.jl). The ecosystem has attracted contributions through software projects at TU Delft, as well as Google Summer of Code and Julia Season of Contributions (in this context, see also Chapter B on supervision engagements).

While Chapter 2 is first and foremost a developer-friendly introduction to our research software package, we include benchmarks of several popular methods for generating CE as part of the exposition of its functionality. The work was presented at JuliaCon Global 2022 and published in proceedings (Altmeyer, Deursen, and Liem 2023a). The chapter makes the following main contributions to the thesis and the field of explainable AI as a whole:

- We fill a gap in the existing open-source software landscape for counterfactual explanations and thus directly address the aspect of human interaction that is needed for trustworthy AI.

- The choice of Julia as a modern, open-source and highly performant programming language, facilitates experimentation with CE methods at an unprecedented scale.
- The vast online documentation accompanying the package and the paper provides an actively maintained, up-to-date introduction not only to our research software, but also the field of CE more generally.
- CounterfactualExplanations.jl has not only powered most of the experiments presented in this thesis, but also external research. It has also laid the foundation for a growing ecosystem of packages geared towards trustworthy AI in Julia.

Chapter 3 is the first traditional research contribution of this thesis. It explores what has been identified in Verma et al. (2022) as one of the core research challenges for the field: the dynamics of recourse. Existing work on CE and AR has largely focused on single individuals in a static environment: given some estimated model, the goal is to find valid counterfactuals for an individual instance that fulfill various desiderata. The ability of such counterfactuals to handle dynamics like data and model drift remains a largely unexplored research challenge. There has also been surprisingly little work on the related question of how the actual implementation of recourse by one individual may affect other individuals. Through this work, we aim to close that gap. We first show that many of the existing methodologies can be collectively described by a generalized framework. We then argue that the existing framework does not account for a hidden external cost of recourse, that only reveals itself when studying the endogenous dynamics of recourse at the group level. Through simulation experiments involving various state-of-the-art counterfactual generators and several benchmark datasets, we generate large numbers of counterfactuals and study the resulting domain and model shifts. We find that the induced shifts are substantial enough to likely impede the applicability of AR in some situations. Fortunately, we find various strategies to mitigate these concerns. Our simulation framework for studying recourse dynamics is fast and open-sourced. This chapter was originally published at the first IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) in 2023 (Altmeyer, Angela, et al. 2023). The key contributions of this work are as follows:

- It demonstrates that long-held beliefs as to what defines optimality in AR, may not always be suitable. Specifically, our experiments show that the application of recourse in practice using off-the-shelf CE methods induces substantial domain and model shifts.
- We argue that these shifts should be considered as a negative externality of individual recourse and call for a paradigm shift from individual to collective recourse in these types of situations.
- By proposing an adapted counterfactual search objective that incorporates this hidden cost, we make that paradigm shift explicit and show that this modified objective lends itself to mitigation strategies.

In recognition of the fact that more plausible counterfactuals are less likely to cause

undesirable dynamics, Chapter 4 explores this desideratum more closely. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic explanations for the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily describe the behavior of the opaque model faithfully. We formalize this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **E**nergy-**C**onstrained **C**onformal **Co**unterfactuals that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that *ECCCo* reconciles the need for faithfulness and plausibility. In particular, we show that for models with gradient access, it is possible to achieve state-of-the-art performance without the need for surrogate models. To do so, our framework relies solely on properties defining the opaque model itself by leveraging recent advances in energy-based modelling and conformal prediction. To our knowledge, this is the first venture in this direction for generating faithful counterfactual explanations. This chapter was originally published at AAAI 2024 (Altmeyer, Farmanbar, et al. 2024a) and makes the following key contributions:

- We show that established measures of model fidelity in XAI in an insufficient evaluation metric for counterfactuals and propose a definition of faithfulness that gives rise to more suitable metrics.
- We introduce *ECCCo*: a novel algorithmic approach aimed at generating energy-constrained conformal counterfactuals that faithfully explain model behavior. We back this claim through extensive empirical evidence demonstrating that *ECCCo* attains plausibility only when appropriate.
- The work lays the foundation for future work aimed at leveraging faithful counterfactuals to improve the trustworthiness of models.

Chapter 5 applies the methods developed in the previous chapter to teach models plausible and actionable explanations. We propose a novel training regime termed counterfactual training that leverages counterfactual explanations to increase the explanatory capacity of models. As discussed above, to be useful in real-word decision-making systems, counterfactuals ought to be (1) plausible with respect to the underlying data and (2) actionable with respect to the user-defined mutability constraints. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. As we demonstrate in Chapter 4, the common objective of developing model-agnostic explainers that deliver plausible explanations for any model is misguided and unnecessary. In Chapter 5, we therefore hold models directly accountable for the desired end goal: counterfactual training employs faithful counterfactuals ad-hoc during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable explanations while promoting robustness and preserving high predictive performance. This work will be published at the 2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

and makes the following key contributions:

- We introduce the methodological framework for counterfactual training (CT) and show theoretically that it can be employed to enforce global actionability constraints.
- Building on previous related work, we propose a new perspective on the link between CE and adversarial examples: specifically, we show and utilize the fact that gradient-based interim ('nascent') CE comply with the standard definition of AE, as samples that have undergone "non-random imperceptible perturbations" (Szegedy et al. 2014).
- Through extensive experiments, we demonstrate that CT substantially improves explainability and positively contributes to the adversarial robustness of trained models without sacrificing predictive performance.

The final research chapter, Chapter 6, explores challenges for trustworthy AI in the age of LLMs. We argue that recent developments in the field of AI, and particularly large language models, have created a 'perfect storm' for observing 'sparks' of Artificial General Intelligence that are spurious. Like simpler models, LLMs distill meaningful representations in their latent embeddings that have been shown to correlate with external variables. Nonetheless, the correlation of such representations has often been linked to human-like intelligence in the latter but not the former. We probe models of varying complexity including random projections, matrix decompositions, deep autoencoders and transformers: all of them successfully distill information that can be used to predict latent or external variables and yet none of them have previously been linked to AGI. We argue and empirically demonstrate that the finding of meaningful patterns in latent spaces of models cannot be seen as evidence in favor of AGI. Additionally, we review literature from the social sciences that shows that humans are prone to seek such patterns and anthropomorphize. We conclude that both the methodological setup and common public image of AI are ideal for the misinterpretation that correlations between model representations and some variables of interest are 'caused' by the model's understanding of underlying 'ground truth' relationships. We, therefore, call for the academic community to exercise extra caution, and to be keenly aware of principles of academic integrity, in interpreting and communicating about AI research outcomes. This work was presented at ECONDAT 2024 and eventually published as a position paper at ICML 2024 (Altmeyer, Demetriou, et al. 2024). We make the following key contributions:

- We present several experiments that may invite claims on models yielding more intelligent outcomes than would have been expected—while at the same time indicating how we feel these claims should *not* be made. Our findings demonstrate that researchers should exert caution when interpreting results from mechanistic interpretability.
- To lend further weight to our argument, we present a review of social science findings in that underline how prone humans are to being enticed by patterns that are not really there.

- We also propose specific structural and cultural changes to improve the current situation by helping researchers avoid common pitfalls.

Finally, we conclude this thesis by discussing the core findings and contributions of this work and proposing directions for future research. To summarize, Table 1.1 provides an overview of the core research chapters along with links to permanent digital object identifiers.

Table 1.1. Overview of replication repositories and DOIs for each chapter.

	Type	DOIs
Chapter 2	package	10.5281/zenodo.8239378; 10.4121/975d2c39-f78e-45d8-a46a-a61e441b1d53
Chapter 3	project	10.5281/zenodo.15309163; 10.4121/d7e7080c-7db1-41e3-95ae-19fa33b4f70c
Chapter 4	package	10.5281/zenodo.15309175; 10.4121/697255aa-c7ad-4bc7-868c-96d00b6aae02
Chapter 5	package	10.5281/zenodo.18374193; 10.4121/3d5a8c54-3c83-4fcb-aea4-f99fafc4edb5
Chapter 6	project	10.5281/zenodo.15309219; 10.4121/d427d182-4bb0-4972-980c-adcb28f430b6

1.8. ORIGINS OF CHAPTERS

This final section of the introduction explains the publication history and author contributions in some more detail. All chapters have undergone thorough peer review and have been published in top-tier academic venues.[3] For all chapters, Patrick Altmeyer was either lead first author or, in the case of Chapter 6, joint first author. Arie van Deursen and Cynthia C. S. Liem have primarily contributed in an editorial capacity consistent with their roles as Patrick's supervisor and daily supervisor, respectively.

Chapter 2 This chapter was published in JuliaCon Proceedings by Patrick Altmeyer, Arie van Deursen and Cynthia C. S. Liem (2023a). The work was presented by Patrick as a main talk at JuliaCon 2022.

Chapter 3 This chapter was published in 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) by Patrick Altmeyer, Giovan Angela,

[3] At the time of printing, Chapter 5 has not yet been published but accepted for publication.

Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen and Cynthia C. S. Liem (2023). Patrick gave an oral and poster presentation at SaTML 2023. Giovan, Aleksander and Karol were all bachelor's students at the time that were co-supervised by Patrick and Cynthia during their final-year research projects (see Appendix B for details).

Chapter 4 This chapter was published in Proceedings of the AAAI Conference on Artificial Intelligence by Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem (2024a). Patrick was joined by Arie to present the work as a poster at AAAI 2024. Mojtaba, who was affiliated with ING Bank at the time, provided expert insights during multiple discussion and editorial meetings.

Chapter 5 This chapter has been accepted for publication at SaTML 2026 and will list Patrick Altmeyer, Aleksander Buszydlik, Arie van Deursen and Cynthia C. S. Liem as authors (2026). Aleksander joined this project at a later stage of the project after finishing his master's degree. He contributed to formal analysis, literature review and writing (both drafting and reviewing), as well as conceptualization, software and visualization for specific evaluation metrics.

Chapter 6 This chapter was published in Proceedings of the 41st International Conference on Machine Learning by Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, Cynthia C. S. Liem (2024). Patrick presented the work as a poster at ICML 2024, but he shared the first-author role with Andrew. Patrick contributed to conceptualization, data curation, formal analysis, investigation, literature review, methodology, project administration, software, visualization and writing (both drafting and reviewing).

2

EXPLAINING BLACK-BOX MODELS THROUGH CounterfactualExplanations.jl

We present CounterfactualExplanations.jl: a package for generating Counterfactual Explanations (CE) and Algorithmic Recourse (AR) for black-box models in Julia. CE explain how inputs into a model need to change to yield specific model predictions. Explanations that involve realistic and actionable changes can be used to provide AR: a set of proposed actions for individuals to change an undesirable outcome for the better. In this article, we discuss the usefulness of CE for Explainable Artificial Intelligence and demonstrate the functionality of our package. The package is straightforward to use and designed with a focus on customization and extensibility. We envision it to one day be the go-to place for explaining arbitrary predictive models in Julia through a diverse suite of counterfactual generators.

This chapter was published in JuliaCon Proceedings by Patrick Altmeyer, Arie van Deursen and Cynthia C. S. Liem (2023a). It provides (1) a gentle introduction to counterfactuals; and (2) an overview of the main open-source Julia package developed as part of this dissertation, and used throughout the thesis to conduct experiments. See Chapter 1.8 for additional publication details.

2.1. INTRODUCTION

Machine Learning models like Deep Neural Networks have become so complex and opaque over recent years that they are generally considered black-box systems. This lack of transparency exacerbates several other problems typically associated with these models: they tend to be unstable (Goodfellow, Shlens, and Szegedy 2015), encode existing biases (Buolamwini and Gebru 2018) and learn representations that are surprising or even counter-intuitive from a human perspective (Buolamwini and Gebru 2018). Nonetheless, they often form the basis for data-driven decision-making systems in real-world applications.

As others have pointed out, this scenario gives rise to an undesirable principal-agent problem involving a group of principals—i.e. human stakeholders—that fail to understand the behavior of their agent—i.e. the black-box system (Borch 2022). The group of principals may include programmers, product managers and other decision-makers who develop and operate the system as well as those individuals ultimately subject to the decisions made by the system. In practice, decisions made by black-box systems are typically left unchallenged since the group of principals cannot scrutinize them:

> "You cannot appeal to (algorithms). They do not listen. Nor do they bend." (O'Neil 2016)

In light of all this, a quickly growing body of literature on Explainable Artificial Intelligence (XAI) has emerged. Counterfactual Explanations fall into this broad category. They can help human stakeholders make sense of the systems they develop, use or endure: they explain how inputs into a system need to change for it to produce different decisions. Explainability benefits internal as well as external quality assurance. Explanations that involve plausible and actionable changes can be used for Algorithmic Recourse (AR): they offer the group of principals a way to not only understand their agent's behavior but also adjust or react to it.

The availability of open-source software to explain black-box models through counterfactuals is still limited. Through the work presented here, we aim to close that gap and thereby contribute to broader community efforts towards XAI. We envision this package to one day be the go-to place for Counterfactual Explanations in Julia.

Thanks to Julia's unique support for interoperability with foreign programming languages we believe that this library may also benefit the broader machine learning and data science community.

Our package provides a simple and intuitive interface to generate CE for many standard classification models trained in Julia, as well as in Python and R. It comes with detailed documentation involving various illustrative example datasets, models and counterfactual generators for binary and multi-class prediction tasks. A carefully designed package architecture allows for a seamless extension of the package functionality through custom generators and models.

The remainder of this article is structured as follows: Section 2.2 presents related work on XAI as well as a brief overview of the methodological framework underlying CE. Section 2.3 introduces the Julia package and its high-level architecture. Section 2.4 presents several basic and advanced usage examples. In Section 2.5 we demonstrate how the package functionality can be customized and extended. To illustrate its practical usability, we explore examples involving real-world data in Section 2.6. Finally, we also discuss the current limitations of our package, as well as its future outlook in Section 2.7. Section 2.8 concludes.

2.2. Background and related work

In this section, we first briefly introduce the broad field of Explainable AI, before narrowing it down to Counterfactual Explanations. We introduce the methodological framework and finally point to existing open-source software.

2.2.1. Literature on Explainable AI

The field of XAI is still relatively young and made up of a variety of subdomains, definitions, concepts and taxonomies. Covering all of these is beyond the scope of this article, so we will focus only on high-level concepts. The following literature surveys provide more detail: Arrieta et al. (2020) provide a broad overview of XAI ; Fan, Xiong, and Wang (2020) focus on explainability in the context of deep learning; and finally, Karimi et al. (2021) and Verma et al. (2022) offer detailed reviews of the literature on Counterfactual Explanations and Algorithmic Recourse (see also Molnar (2022) and Varshney (2022)). T. Miller (2019) explicitly discusses the concept of explainability from the perspective of a social scientist.

The first broad distinction we want to make here is between **Interpretable** and **Explainable** AI. These terms are often used interchangeably, but this can lead to confusion. We find the distinction made in Rudin (2019) useful: Interpretable AI involves models that are inherently interpretable and transparent such as general additive models (GAM), decision trees and rule-based models; Explainable AI involves models that are not inherently interpretable but require additional tools to be explainable to humans. Examples of the latter include Ensembles, Support Vector

Machines and Deep Neural Networks. Some would argue that we best avoid the second category of models altogether and instead focus solely on interpretable AI Rudin (2019). While we agree that initial efforts should always be geared towards interpretable models, avoiding black boxes altogether would entail missed opportunities and anyway is probably not very realistic at this point. For that reason, we expect the need for XAI to persist in the medium term. Explainable AI can further be broadly divided into **global** and **local** explainability: the former is concerned with explaining the average behavior of a model, while the latter involves explanations for individual predictions (Molnar 2022). Tools for global explainability include partial dependence plots (PDP), which involve the computation of marginal effects through Monte Carlo, and global surrogates. A surrogate model is an interpretable model that is trained to explain the predictions of a black-box model.

Counterfactual Explanations fall into the category of local methods: they explain how individual predictions change in response to individual feature perturbations. Among the most popular alternatives to Counterfactual Explanations are local surrogate explainers including Local Interpretable Model-agnostic Explanations (LIME) and Shapley additive explanations (SHAP). Since explanations produced by LIME and SHAP typically involve simple feature importance plots, they arguably rely on reasonably interpretable features at the very least. Contrary to Counterfactual Explanations, for example, it is not obvious how to apply LIME and SHAP to high-dimensional image data. Nonetheless, local surrogate explainers are among the most widely used XAI tools today, potentially because they are easy to interpret and implemented in popular programming languages. Proponents of surrogate explainers also commonly mention that there is a straightforward way to assess their reliability: a surrogate model that generates predictions in line with those produced by the black-box model is said to have high **fidelity** and therefore considered reliable. As intuitive as this notion may be, it also points to an obvious shortfall of surrogate explainers: even a high-fidelity surrogate model that produces the same predictions as the black-box model 99 per cent of the time is useless and potentially misleading for every 1 out of 100 individual predictions.

A recent study has shown that even experienced data scientists tend to put too much trust in explanations produced by LIME and SHAP (Kaur et al. 2020). Another recent work has shown that both methods can be easily fooled: they depend on random input perturbations, a property that can be abused by adverse agents to essentially whitewash strongly biased black-box models (Slack et al. 2020). In related work, the same authors find that while gradient-based Counterfactual Explanations can also be manipulated, there is a straightforward way to protect against this in practice (Slack et al. 2021). In the context of quality assessment, it is also worth noting that—contrary to surrogate explainers—CE always achieve full fidelity by construction: counterfactuals are searched with respect to the black-box classifier, not some proxy for it. That being said, CE should also be used with care and research around them is still in its early stages.

2.2.2. A framework for counterfactual explanations

Counterfactual search involves feature perturbations: we are interested in understanding how we need to change individual attributes in order to change the model output to a desired value or label (Molnar 2022). Typically, the underlying methodology is presented in the context of binary classification: $M : \mathcal{X} \mapsto \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^D$ and $\mathcal{Y} = \{0, 1\}$. Further, let $t = 1$ be the target class and let x denote the factual feature vector of some individual sample outside the target class, so $y = M(x) = 0$. We follow this convention here, though it should be noted that the ideas presented here also carry over to multi-class problems and regression (Molnar 2022).

The counterfactual search objective originally proposed by Wachter, Mittelstadt, and Russell (2017) is as follows

$$\min_{x' \in \mathcal{X}} h(x') \quad \text{s. t.} \quad M(x') = t \tag{2.1}$$

where $h(\cdot)$ quantifies how complex or costly it is to go from the factual x to the counterfactual x'. To simplify things we can restate this constrained objective as the following unconstrained and differentiable problem:

$$x' = \arg\min_{x'} \ell(M(x'), t) + \lambda h(x') \tag{2.2}$$

Here ℓ denotes some loss function targeting the deviation between the target label and the predicted label and λ governs the strength of the complexity penalty. Provided we have gradient access for the black-box model M the solution to this problem can be found through gradient descent. This generic framework lays the foundation for most state-of-the-art approaches to counterfactual search and is also used as the baseline approach in our package. The hyperparameter λ is typically tuned through grid search or in some sense pre-determined by the nature of the problem. Conventional choices for ℓ include margin-based losses like cross-entropy loss and hinge loss. It is worth pointing out that the loss function is typically computed with respect to logits rather than predicted probabilities, a convention that we have chosen to follow.[1]

Numerous extensions to this simple approach have been developed since CE were first proposed in 2017 (see Verma et al. (2022) and Karimi et al. (2021) for surveys). The various approaches largely differ in that they use different flavors of search objective defined in Equation 3.2. Different penalties are often used to address many of the desirable properties of effective CE that have been set out. These desiderata include: **proximity** — the distance between factual and counterfactual features should be small (Wachter, Mittelstadt, and Russell 2017); **actionability** —

[1] Implementations of loss functions with respect to logits are often numerically more stable. For example, the `logitbinarycrossentropy(ŷ, y)` implementation in `Flux.Losses` (used here) is more stable than the mathematically equivalent `binarycrossentropy(ŷ, y)`.

the proposed recourse should be actionable (Ustun, Spangher, and Liu 2019; Poyiadzi et al. 2020); **plausibility** — the counterfactual explanation should be plausible to a human (Joshi et al. 2019; Schut et al. 2021); **sparsity** — the counterfactual explanation should involve as few individual feature changes as possible (Schut et al. 2021); **robustness** — the counterfactual explanation should be robust to domain and model shifts (Upadhyay, Joshi, and Lakkaraju 2021); **diversity** — ideally multiple diverse counterfactuals should be provided (Mothilal, Sharma, and Tan 2020); and **causality** — counterfactuals should respect the structural causal model underlying the data generating process (Karimi et al. 2020; Karimi, Schölkopf, and Valera 2021).

Beyond gradient-based counterfactual search, which has been the main focus in our development so far, various methodologies have been proposed that can handle non-differentiable models like decision trees. We have implemented some of these approaches and will discuss them further in Section 2.3.2.

2.2.3. EXISTING SOFTWARE

To the best of our knowledge, the package introduced here provides the first implementation of Counterfactual Explanations in Julia and therefore represents a novel contribution to the community. As for other programming languages, we are only aware of one other unifying framework: the Python library CARLA (Pawelczyk et al. 2021).[2] In addition to that, there exists open-source code for some specific approaches to CE that have been proposed in recent years. The approach-specific implementations that we have been able to find are generally well-documented, but exclusively in Python. For example, a PyTorch implementation of a greedy generator for Bayesian models proposed in Schut et al. (2021) has been released. As another example, the popular InterpretML library includes an implementation of a diverse counterfactual generator (Mothilal, Sharma, and Tan 2020).

Generally speaking, software development in the space of XAI has largely focused on various global methods and surrogate explainers: implementations of PDP, LIME and SHAP are available for both Python (e.g. `lime`, `shap`) and R (e.g. `lime`, `iml`, `shapper`, `fastshap`). In the Julia space, there exist two packages related to XAI: firstly, `ShapML.jl`, which provides a fast implementation of SHAP; and, secondly, `ExplainableAI.jl`, which enables users to easily visualise gradients and activation maps for `Flux.jl` models. We also should not fail to mention the comprehensive Interpretable AI infrastructure, which focuses exclusively on interpretable models.

Arguably the current availability of tools for explaining black-box models in Julia is limited, but it appears that the community is invested in changing that. The team behind `MLJ.jl`, for example, recruited contributors for a project about both

[2] While we were writing this paper, the `R` package `counterfactuals` was released (Dandl et al. 2023). The developers seem to also envision a unifying framework, but the project appears to still be in its early stages.

Interpretable and Explainable AI in 2022.[3] With our work on Counterfactual Explanations we hope to contribute to these efforts. We think that because of its unique transparency the Julia language naturally lends itself towards building Trustworthy AI systems.

2.3. INTRODUCING: COUNTERFACTUALEXPLANATIONS.JL

Figure 2.1 provides an overview of the package architecture. It is built around two core modules that are designed to be as extensible as possible through dispatch: 1) `Models` is concerned with making any arbitrary model compatible with the package; 2) `Generators` is used to implement counterfactual search algorithms. The core function of the package—`generate_counterfactual`—uses an instance of type `<:AbstractFittedModel` produced by the `Models` module and an instance of type `<:AbstractGenerator` produced by the `Generators` module. Relating this to the methodology outlined in Section 2.2.2, the former instance corresponds to the model M, while the latter defines the rules for the counterfactual search (Equation 3.2).

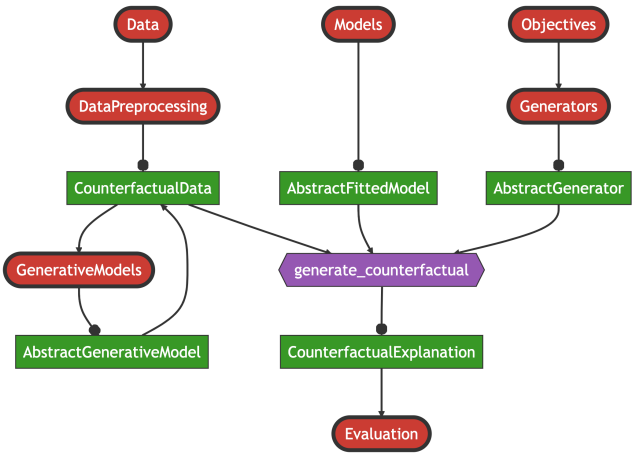


Figure 2.1. High-level schematic overview of package architecture. Modules are shown in red, structs in green and functions in purple.

2.3.1. MODELS

The package currently offers native support for models built and trained in Flux (Mike Innes 2018) as well as a small subset of models made available through MLJ

[3]For details, see the Google Summer of Code 2022 project proposal: https://julialang.org/jsoc/gsoc/MLJ/#interpretable_machine_learning_in_julia.

(Blaom et al. 2020). While in general it is assumed that users resort to this package to explain their pre-trained models, we provide a simple API call to train the following models:

- Linear Classifier (Logistic Regression and Multinomial Logit)
- Multi-Layer Perceptron (Deep Neural Network)
- Deep Ensemble Lakshminarayanan, Pritzel, and Blundell (2017)
- Decision Tree, Random Forest, Gradient Boosted Trees

As we demonstrate below, it is straightforward to extend the package through custom models. Support for `torch` models trained in Python or R is also available.[4]

2.3.2. GENERATORS

A large and growing number of counterfactual generators have already been implemented in our package (Table 2.1). At a high level, we distinguish generators in terms of their compatible model types, their default search space, and their composability. All "gradient-based" generators are compatible with differentiable models, e.g. `Flux` and `torch`, while "tree-based" generators are only applicable to models that involve decision trees. Concerning the search space, it is possible to search counterfactuals in a lower-dimensional latent embedding of the feature space that implicitly encodes the data-generating process (DGP). To learn the latent embedding, existing work has typically relied on generative models or existing causal knowledge (Joshi et al. 2019; Karimi, Schölkopf, and Valera 2021). While this notion is compatible with all of our gradient-based generators, only some generators search a latent space by default. Finally, composability implies that the given generator can be blended with any other composable generator, which we discuss in Section 2.4.2.

Beyond these broad technical distinctions, generators largely differ in terms of how they address the various desiderata mentioned above: *ClapROAR* aims to preserve the classifier, i.e. to generate counterfactuals that are robust to endogenous model shifts (Chapter 3); *CLUE* searches plausible counterfactuals in the latent embedding of a generative model by explicitly minimizing predictive entropy (Antorán et al. 2020); *DiCE* is designed to generate multiple, maximally diverse counterfactuals (Mothilal, Sharma, and Tan 2020); *FeatureTweak* leverages the internals of decision trees to search counterfactuals on a feature-by-feature basis, finding the counterfactual that tweaks the features in the least costly way (Tolomei et al. 2017); *Gravitational* aims to generate plausible and robust counterfactuals by minimizing the distance to observed samples in the target class (Chapter 3); *Greedy* aims to generate plausible counterfactuals by implicitly minimizing predictive uncertainty of Bayesian classifiers (Schut et al. 2021); *GrowingSpheres* is model-agnostic, relying solely on identifying nearest neighbors of counterfactuals in the target class by

[4] We are currently relying on `PythonCall.jl` and `RCall.jl` and this functionality is still somewhat brittle. Since this is more of an edge case, we may move this feature into its own package in the future.

gradually increasing the search radius and then moving counterfactuals in that direction (Laugel et al. 2017); *PROBE* generates probabilistically robust counterfactuals (Pawelczyk et al. 2023); *REVISE* addresses the need for plausibility by searching counterfactuals in the latent embedding of a Variational Autoencoder (VAE) (Joshi et al. 2019); *Wachter* is the baseline approach that only penalizes the distance to the original sample (Wachter, Mittelstadt, and Russell 2017).

Table 2.1. Overview of implemented counterfactual generators.

Generator	Model Type	Search Space	Composable
ClaPROAR (Altmeyer, Angela, et al. 2023)	gradient based	feature	yes
CLUE (Antorán et al. 2020)	gradient based	latent	yes
DiCE (Mothilal, Sharma, and Tan 2020)	gradient based	feature	yes
FeatureTweak (Tolomei et al. 2017)	tree based	feature	no
Gravitational (Altmeyer, Angela, et al. 2023)	gradient based	feature	yes
Greedy (Schut et al. 2021)	gradient based	feature	yes
GrowingSpheres (Laugel et al. 2017)	agnostic	feature	no
PROBE (Pawelczyk et al. 2023)	gradient based	feature	no
REVISE (Joshi et al. 2019)	gradient based	latent	yes
Wachter (Wachter, Mittelstadt, and Russell 2017)	gradient based	feature	yes

2.3.3. DATA CATALOGUE

To allow researchers and practitioners to test and compare counterfactual generators, the package ships with catalogues of pre-processed synthetic and real-world benchmark datasets from different domains. Real-world datasets include:

- Adult Census (Barry Becker 1996)
- California Housing (Pace and Barry 1997)
- CIFAR10 (Krizhevsky 2009)
- German Credit (Hoffman 1994)
- Give Me Some Credit (Kaggle 2011)
- MNIST (LeCun et al. 1998) and Fashion MNIST (Xiao, Rasul, and Vollgraf 2017)
- UCI defaultCredit (Yeh and Lien 2009)

Custom datasets can also be easily preprocessed as explained in the documentation.

2.3.4. PLOTTING

The package also extends common `Plots.jl` methods to facilitate the visualization of results. Calling the generic `plot()` method on an instance of type `<:CounterfactualExplanation`, for example, generates a plot visualizing the entire counterfactual path in the feature space[5]. We will see several examples of this below.

2.4. BASIC USAGE

In the following, we begin our exploration of the package functionality with a simple example. We then demonstrate how more advanced generators can be easily composed and show how users can impose mutability constraints on features. Finally, we also briefly explore the topics of counterfactual evaluation and benchmarking.

2.4.1. A SIMPLE GENERIC GENERATOR

Listing 2.1 provides a complete example demonstrating how the framework presented in Section 2.2.2 can be implemented through our package: using a synthetic data set with linearly separable features we first fit a linear classifier; next, we define the target class and then draw a random sample from the other class; finally, we instantiate a generic generator and run the counterfactual search. Figure 2.2 illustrates the resulting counterfactual path in the two-dimensional feature space. Features go through iterative perturbations until the desired confidence level is reached as illustrated by the contour in the background, which shows the softmax output for the target class.

[5] For multidimensional input data, standard dimensionality reduction techniques are used to compress the data. In this case, the classifier's decision boundary is approximated through a Nearest neighbor model. This is still somewhat experimental and will be improved in the future.

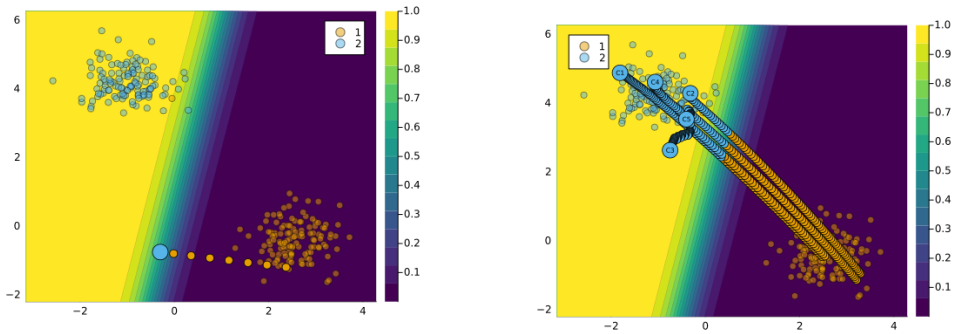
Listing 2.1 Standard workflow for generating counterfactuals.

```
# Data and Classifier:
counterfactual_data = load_linearly_separable()                    ①
M = fit_model(counterfactual_data, :Linear)

# Factual and Target:
yhat = predict_label(M, counterfactual_data)
target = 2                                                         ②
candidates = findall(vec(yhat) .!= target)
chosen = rand(candidates)
x = select_factual(counterfactual_data, chosen)                    ③

# Counterfactual search:
generator = GenericGenerator()                                     ④
ce = generate_counterfactual(                                      ⑤
    x, target, counterfactual_data, M, generator)
```

① Load synthetic data and fit linear model.
② Define the target class.
③ Draw a random sample from the other class.
④ Instantiate a generic generator.
⑤ Run the counterfactual search.



(a) Counterfactual path using generic counterfactual generator for conventional binary classifier.

(b) Counterfactual path using the *DiCE* generator.

Figure 2.2. Counterfactual explanations for a binary classifier.

In this simple example, the generic generator produces a valid counterfactual, since the decision boundary is crossed and the predicted label is flipped. But the counterfactual is not plausible: it does not appear to be generated by the same DGP

as the observed data in the target class. This is because the generic generator does not take into account any of the desiderata mentioned in Section 2.2.2, except for the distance to the factual sample.

2.4.2. COMPOSING GENERATORS

To address these issues, we can leverage the ideas underlying some of the more advanced counterfactual generators introduced above. In particular, we will now show how easy it is to compose custom generators that blend different ideas through user-friendly macros.

Suppose we wanted to address the desiderata of plausibility and diversity. We could do this by blending ideas underlying the *DiCE* generator with the *REVISE* generator. Formally, the corresponding search objective would be defined as follows,

$$\mathbf{Z}' = \arg\min_{\mathbf{Z}' \in \mathcal{Z}^{L \times K}} \{\ell(M(f(\mathbf{Z}')), t) + \lambda \cdot \text{diversity}(f(\mathbf{Z}'))\} \tag{2.3}$$

where $\mathbf{X}'$ is an L-dimensional array of counterfactuals, $f : \mathcal{Z}^{L \times K} \mapsto \mathcal{X}^{L \times D}$ is a function that maps the $L \times K$-dimensional latent space $\mathcal{Z}$ to the $L \times D$-dimensional feature space $\mathcal{X}$ and diversity$(\cdot)$ is the penalty proposed by Mothilal, Sharma, and Tan (2020) that induces diverse sets of counterfactuals. As in Equation 3.2, ℓ is the loss function, M is the black-box model, t is the target class, and λ is the strength of the penalty.

Listing 2.2 demonstrates how Equation 2.3 can be seamlessly translated into Julia code. We begin by instantiating a `GradientBasedGenerator`. Next, we use chained macros for composition: firstly, we define the counterfactual search `@objective` corresponding to *DiCE* (Mothilal, Sharma, and Tan 2020); secondly, we define the latent space search strategy corresponding to *REVISE* (Joshi et al. 2019) using the `@search_latent_space` macro; finally, we specify our preferred optimization method using the `@with_optimiser` macro.

In this case, the counterfactual search is performed in the latent space of a Variational Autoencoder (VAE) that is automatically trained on the observed data. It is important to specify the keyword argument `num_counterfactuals` of the `generate_counterfactual` to some value higher than 1 (default), to ensure that the diversity penalty is effective. The resulting counterfactual path is shown in Figure 2.2b below. We observe that the resulting counterfactuals are diverse and the majority of them are plausible.

2.4.3. MUTABILITY CONSTRAINTS

In practice, features usually cannot be perturbed arbitrarily. Suppose, for example, that one of the features used by a bank to predict the creditworthiness of its clients is *age*. If a counterfactual explanation for the prediction model indicates that older

Listing 2.2 Composing a custom generator.

```
# Composition:
generator = GradientBasedGenerator()                    ①
@chain generator begin
    @objective logitcrossentropy                        ②
        + 0.2ddp_diversity
    @search_latent_space                                ③
    @with_optimiser Adam(0.005)                         ④
end
```

① Instantiate a `GradientBasedGenerator`.
② Define the counterfactual search `@objective` corresponding to *DiCE* (Mothilal, Sharma, and Tan 2020).
③ Define the latent space search strategy corresponding to *REVISE* (Joshi et al. 2019).
④ Specify optimization method.

clients should "grow younger" to improve their creditworthiness, then this is an interesting insight (it reveals age bias), but the provided recourse is not actionable. In such cases, we may want to constrain the mutability of features. To illustrate how this can be implemented in our package, we will continue with the example from above.

Mutability can be defined in terms of four different options: 1) the feature is mutable in both directions, 2) the feature can only increase (e.g. *age*), 3) the feature can only decrease (e.g. *time left* until your next deadline) and 4) the feature is not mutable (e.g. *skin colour*, *ethnicity*, …). To specify which category a feature belongs to, users can pass a vector of symbols containing the mutability constraints at the preprocessing stage. For each feature one can choose from these four options: `:both` (mutable in both directions), `:increase` (only up), `:decrease` (only down) and `:none` (immutable). By default, `nothing` is passed to that keyword argument and it is assumed that all features are mutable in both directions.[6]

We can impose that the first feature is immutable as follows:

```
counterfactual_data.mutability = [:none, :both]
```

The resulting counterfactual path is shown in Figure 2.3 below. Since only the second feature can be perturbed, the sample can only move along the vertical axis. In this case, the counterfactual search does not yield a valid counterfactual, since the target class is not reached.

[6] Mutability constraints are not yet implemented for Latent Space search.

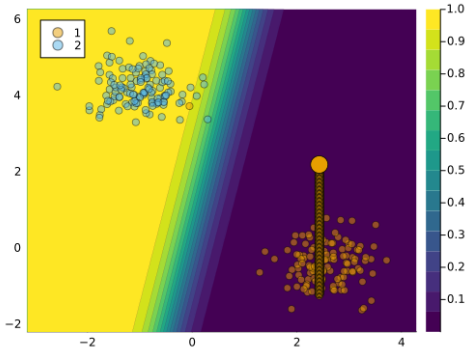


Figure 2.3. Counterfactual path with immutable feature.

2.4.4. EVALUATION AND BENCHMARKING

The package also makes it easy to evaluate counterfactuals with respect to many of the desiderata mentioned above. For example, users may want to infer how costly the provided recourse is to individuals. To this end, we can measure the distance of the counterfactual from its original value. The API call to compute the distance metric defined in Wachter, Mittelstadt, and Russell (2017), for instance, is as simple as `evaluate(ce; measure=distance_mad)`, where `ce` can also be a vector of `CounterfactualExplanations`.

Additionally, the package provides a benchmarking framework that allows users to compare the performance of different generators on a given dataset. In Figure 2.4 we show the results of a benchmark comparing several generators in terms of the average cost and implausibility of the generated counterfactuals. The cost is proxied by the L1-norm of the difference between the factual and counterfactual features, while implausibility is measured by the distance of the counterfactuals from samples in the target class. The results illustrate that there is a tradeoff between minimizing costs to individuals and generating plausible counterfactuals.

2.5. CUSTOMIZATION AND EXTENSIBILITY

One of our priorities has been to make our package customizable and extensible. In the long term, we aim to add support for more default models and counterfactual generators. In the short term, it is designed to allow users to integrate models and generators themselves. These community efforts will facilitate our long-term goals.

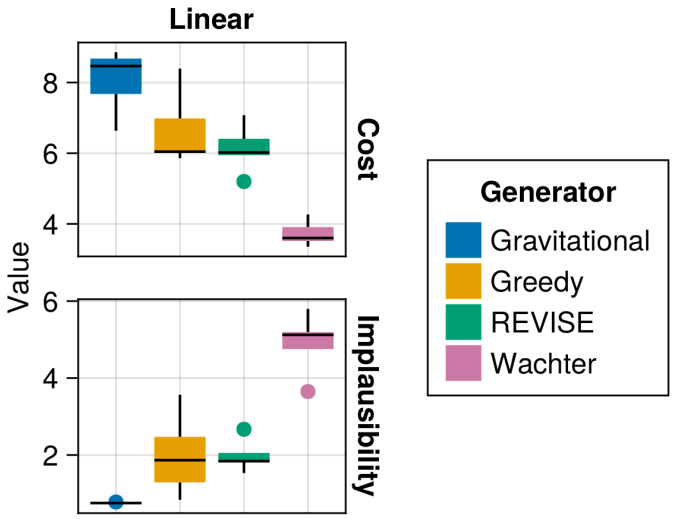


Figure 2.4. Benchmarking results for different generators.

2.5.1. ADDING CUSTOM MODELS

At the high level, only two steps are necessary to make any supervised learning model compatible with our package:

- **Subtyping**: We need to subtype the `AbstractFittedModel`.
- **Dispatch**: The functions`logits` and `probs` need to be extended through custom methods for the model in question.

To demonstrate how this can be done in practice, we will reiterate here how native support for `Flux.jl` (Mike Innes 2018) deep learning models was enabled.[7] Once again we use synthetic data for an illustrative example. Listing 2.3 below builds a simple model architecture that can be used for a multi-class prediction task. Note how outputs from the final layer are not passed through a softmax activation function, since the counterfactual loss is evaluated with respect to logits as we discussed earlier. The model is trained with dropout.

Listing 2.4 implements the two steps that were necessary to make Flux models compatible with the package. First, we declare our new struct as a subtype of `AbstractDifferentiableModel`, which itself is an abstract subtype of `AbstractFittedModel`.[8] Computing logits amounts to just calling the model on

[7] Flux models are now natively supported by our package and can be instantiated by calling `FluxModel()`.

[8] Note that we also provide a field determining the likelihood. This is optional and only used internally to determine which loss function to use in the counterfactual search. If this field is not provided to the model, the loss function needs to be explicitly supplied to the generator.

Listing 2.3 A simple neural network model.

```
n_hidden = 32
output_dim = size(y,1)
input_dim = 2
activation =
model = Flux.Chain(
    Dense(input_dim, n_hidden, activation),
    Dropout(0.1),
    Dense(n_hidden, output_dim)
)
```

Listing 2.4 A wrapper for Flux models.

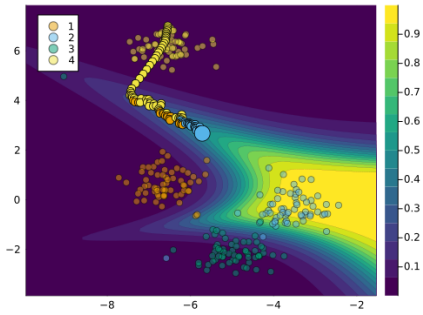
```
# Step 1)
struct MyFluxModel <: AbstractDifferentiableModel                    ①
    model::Any
    likelihood::Symbol                                               ②
end

# Step 2)
# import functions in order to extend
import CounterfactualExplanations.Models: logits
import CounterfactualExplanations.Models: probs
logits(M::MyFluxModel, X::AbstractArray) = M.model(X)
probs(M::MyFluxModel, X::AbstractArray) = softmax(logits(M, X))
M = MyFluxModel(model, :classification_multi)                        ③
```

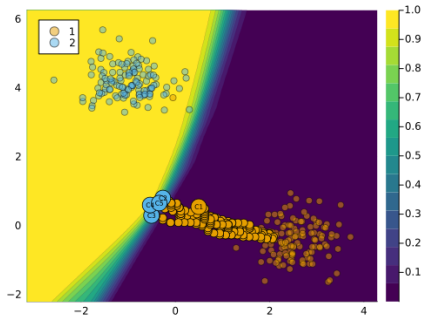
① Declare new struct as a subtype of `AbstractDifferentiableModel`.
② Optional field specifying the likelihood (used internally to guess loss function).
③ Instantiate custom model by wrapping the neural network (Listing 2.3) and specifying likelihood.

inputs. Predicted probabilities for labels can be computed by passing logits through the softmax function.

The API call for generating counterfactuals for our new model is the same as before. Figure 2.5a shows the resulting counterfactual path for a randomly chosen sample. In this case, the contour shows the predicted probability that the input is in the target class ($t = 2$).



(a) Counterfactual path using generic counterfactual generator for multi-class classifier.



(b) Counterfactual path for a generator with dropout.

Figure 2.5. Counterfactual explanations for custom models and generators.

2.5.2. ADDING CUSTOM GENERATORS

In some cases, composability may not be sufficient to implement specific logics underlying certain counterfactual generators. In such cases, users may want to implement custom generators. To illustrate how this can be done we will consider a simple extension of our `GenericGenerator`. As we have seen above, Counterfactual Explanations are not unique. In light of this, we might be interested in quantifying the uncertainty around the generated counterfactuals (Delaney, Greene, and Keane 2021). One idea could be, to use dropout to randomly switch features on and off in each iteration. Without dwelling further on the merit of this idea, we will now briefly show how this can be implemented.

2.5.2.1. A GENERATOR WITH DROPOUT

Listing 2.5 implements two important steps: 1) create an abstract subtype of the `AbstractGradientBasedGenerator` and 2) create a constructor with an additional field for the dropout probability.

Next, in Listing 2.6 we define how feature perturbations are generated for our custom dropout generator: in particular, we extend the relevant function through a method that implements the dropout logic.

Finally, we proceed to generate counterfactuals in the same way we always do. The resulting counterfactual path is shown in Figure 2.5b.

2

Listing 2.5 Building a custom generator with dropout.

```julia
# Abstract suptype:
abstract type
    AbstractDropoutGenerator <: AbstractGradientBasedGenerator
end                                                                    ①

# Constructor:
struct DropoutGenerator <: AbstractDropoutGenerator                    ②
    loss::Function # loss function
    penalty::Function
     ::AbstractFloat # strength of penalty
    latent_space::Bool
    opt::Any # optimizer
    p_dropout::AbstractFloat # dropout rate
end
```

① Create an abstract subtype of the `AbstractGradientBasedGenerator`.
② Create a constructor with an additional field for the dropout probability.

Listing 2.6 Generating feature perturbations with dropout.

```julia
using CounterfactualExplanations.Generators
using StatsBase
function Generators.generate_perturbations(
    generator::AbstractDropoutGenerator,
    ce::CounterfactualExplanation
)
    s  = deepcopy(ce.s )
    new_s  = Generators.propose_state(generator, ce)
    Δs  = new_s  - s  # gradient step

    # Dropout:
    set_to_zero = sample(
        1:length(Δs ),
        Int(round(generator.p_dropout*length(Δs ))),
        replace=false
    )
    Δs [set_to_zero] .= 0
    return Δs
end
```

2.6. REAL-WORLD EXAMPLES

Now that we have explained the basic functionality of `CounterfactualExplanations.jl` through some synthetic examples, it is time to work through examples involving real-world data.

2.6.1. GIVE ME SOME CREDIT

The *Give Me Some Credit* dataset is one of the tabular real-world datasets that ship with the package (Kaggle 2011). It can be used to train a binary classifier to predict whether a borrower is likely to experience financial difficulties in the next two years. In particular, we have an output variable $y \in \{0 = \texttt{no stress}, 1 = \texttt{stress}\}$ and a feature matrix X that includes socio-demographic variables like `age` and `income`. A retail bank might use such a classifier to determine if potential borrowers should receive credit or not.

For the classification task, we use a Multi-Layer Perceptron with dropout regularization. Using the Gravitational generator (Chapter 3) we will generate counterfactuals for ten randomly chosen individuals that would be denied credit based on our pre-trained model. Concerning the mutability of features, we only impose that the `age` cannot be decreased.

Figure 2.6 shows the resulting counterfactuals proposed by Wachter in the two-dimensional feature space spanned by the `age` and `income` variables. An increase in income and age is recommended for the majority of individuals, which seems plausible: both age and income are typically positively related to creditworthiness.

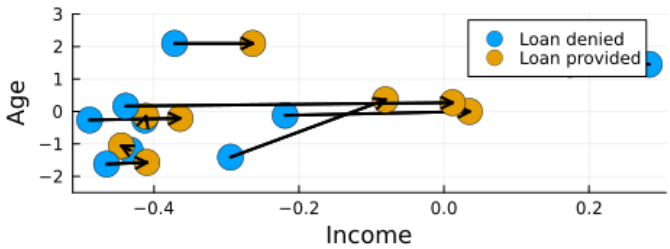


Figure 2.6. Give Me Some Credit: counterfactuals for would-be borrowers proposed by the Gravitational Generator.

2.6.2. MNIST

For our second example, we will look at image data. The MNIST dataset contains 60,000 training samples of handwritten digits in the form of 28x28 pixel grey-scale images (LeCun et al. 1998). Each image is associated with a label indicating the

digit (0-9) that the image represents. The data makes for an interesting case study of CE because humans have a good idea of what plausible counterfactuals of digits look like. For example, if you were asked to pick up an eraser and turn the digit in the left panel of Figure 2.7 into a four (4) you would know exactly what to do: just erase the top part.

Listing 2.7 Loading pre-trained models and data for MNIST.

```
counterfactual_data = load_mnist()
X, y = unpack_data(counterfactual_data)
input_dim, n_obs = size(counterfactual_data.X)
M = load_mnist_mlp()
vae = load_mnist_vae()
vae_weak = load_mnist_vae(;strong=false)
```

On the model side, we will use a simple multi-layer perceptron (MLP). Listing 2.7 loads the data and the pre-trained MLP. It also loads two pre-trained Variational Auto-Encoders, which will be used by our counterfactual generator of choice for this task: *REVISE*.

The proposed counterfactuals are shown in Figure 2.7. In the case in which *REVISE* has access to an expressive VAE (centre), the result looks convincing: the perturbed image does look like it represents a four (4). In terms of explainability, we may conclude that removing the top part of the handwritten nine (9) leads the black-box model to predict that the perturbed image represents a four (4). We should note, however, that the quality of counterfactuals produced by *REVISE* hinges on the performance of the underlying generative model, as demonstrated by the result on the right. In this case, *REVISE* uses a weak VAE and the resulting counterfactual is invalid. In light of this, we recommend using Latent Space search with care.

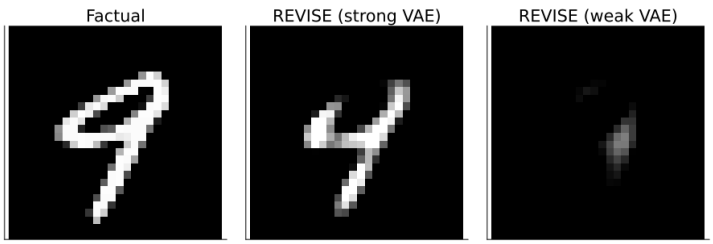


Figure 2.7. Counterfactual explanations for MNIST using a Latent Space generator: turning a nine (9) into a four (4).

2.7. DISCUSSION AND OUTLOOK

We believe that this package in its current form offers a valuable contribution to ongoing efforts towards XAI in Julia. That being said, there is significant scope for future developments, which we briefly outline in this final section.

2.7.1. CANDIDATE MODELS AND GENERATORS

The package supports various models and generators either natively or through minimal augmentation. In future work, we would like to prioritize the addition of further predictive models and generators. Concerning the former, it would be useful to add native support for any supervised models built in `MLJ.jl`, an extensive Machine Learning framework for Julia (Blaom et al. 2020). This may also involve adding support for regression models as well as additional non-differentiable models. In terms of counterfactual generators, there is a list of recent methodologies that we would like to implement including MINT (Karimi, Schölkopf, and Valera 2021), ROAR (Upadhyay, Joshi, and Lakkaraju 2021) and FACE (Poyiadzi et al. 2020).

2.7.2. ADDITIONAL DATASETS

For benchmarking and testing purposes it will be crucial to add more datasets to our library. We have so far prioritized tabular datasets that have typically been used in the literature on counterfactual explanations including *Adult*, *Give Me Some Credit* and *German Credit* (Karimi et al. 2021). There is scope for adding data sources that have so far not been explored much in this context including additional image datasets as well as audio, natural language and time-series data.

2.8. CONCLUDING REMARKS

`CounterfactualExplanation.jl` is a package for generating Counterfactual Explanations and Algorithmic Recourse in Julia. Through various synthetic and real-world examples, we have demonstrated the basic usage of the package as well as its extensibility. The package has already served us in our research to benchmark various methodological approaches to Counterfactual Explanations and Algorithmic Recourse. We therefore strongly believe that it should help other practitioners and researchers in their own efforts towards Trustworthy AI.

We envision this package to one day constitute the go-to place for explaining arbitrary predictive models through an extensive suite of counterfactual generators. As a major next step, we aim to make our library as compatible as possible with the popular `MLJ.jl` package for machine learning in Julia. We invite the Julia community to contribute to these goals through usage, open challenge and active development.

2.9. ACKNOWLEDGEMENTS

We are immensely grateful to the group of TU Delft students who contributed huge improvements to this package as part of a university project in 2023: Rauno Arike, Simon Kasdorp, Lauri Keskül, Mariusz Kicior, Vincent Pikand. We also want to thank the broader Julia community for being welcoming and open and for supporting research contributions like this one. Some of the members of TU Delft were partially funded by ICAI AI for Fintech Research, an ING—TU Delft collaboration.

3

ENDOGENOUS MACRODYNAMICS IN ALGORITHMIC RECOURSE

Existing work on Counterfactual Explanations (CE) and Algorithmic Recourse (AR) has largely focused on single individuals in a static environment: given some estimated model, the goal is to find valid counterfactuals for an individual instance that fulfill various desiderata. The ability of such counterfactuals to handle dynamics like data and model drift remains a largely unexplored research challenge. There has also been surprisingly little work on the related question of how the actual implementation of recourse by one individual may affect other individuals. Through this work, we aim to close that gap. We first show that many of the existing methodologies can be collectively described by a generalized framework. We then argue that the existing framework does not account for a hidden external cost of recourse, that only reveals itself when studying the endogenous dynamics of recourse at the group level. Through simulation experiments involving various state-of-the-art counterfactual generators and several benchmark datasets, we generate large numbers of counterfactuals and study the resulting domain and model shifts. We find that the induced shifts are substantial enough to likely impede the applicability of Algorithmic Recourse in some situations. Fortunately, we find various strategies to mitigate these concerns. Our simulation framework for studying recourse dynamics is fast and open-sourced.

This chapter was published in 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) by Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen and Cynthia C. S. Liem (2023). See Chapter 1.8 for additional publication details.

3.1. INTRODUCTION

Recent advances in Artificial Intelligence (AI) have propelled its adoption in scientific domains outside of Computer Science including Healthcare, Bioinformatics, Genetics and the Social Sciences. While this has in many cases brought benefits in terms of efficiency, state-of-the-art models like Deep Neural Networks (DNN) have also given rise to a new type of problem in the context of data-driven decision-making. They are essentially **black boxes**: so complex, opaque and underspecified in the data that it is often impossible to understand how they actually arrive at their decisions without auxiliary tools. Despite this shortcoming, black-box models have grown in popularity in recent years and have at times created undesirable societal outcomes (O'Neil 2016). The scientific community has tackled this issue from two different angles: while some have appealed for a strict focus on inherently interpretable models (Rudin 2019), others have investigated different ways to explain the behavior of black-box models. These two subdomains can be broadly referred to as **interpretable AI** and **explainable AI** (XAI), respectively.

Among the approaches to XAI that have recently grown in popularity are **Counterfactual Explanations** (CE). They explain how inputs into a model need to change for it to produce different outputs. Counterfactual Explanations that involve realistic and actionable changes can be used for the purpose of **Algorithmic Recourse** (AR) to help individuals who face adverse outcomes. An example relevant to the Social Sciences is consumer credit: in this context, AR can be used to guide individuals in improving their creditworthiness, should they have previously been denied access to credit based on an automated decision-making system. A meaningful recourse recommendation for a denied applicant could be: *"If your net savings rate had been 10% of your monthly income instead of the actual 8%, your application would have been successful. See if you can temporarily cut down on consumption."* In the remainder of this paper, we will use both terminologies—recourse and counterfactual—interchangeably to refer to situations where counterfactuals are generated with the intent to provide individual recourse.

Existing work in this field has largely worked in a static setting: various approaches have been proposed to generate counterfactuals for a given individual that is subject to some pre-trained model. More recent work has compared different approaches within this static setting (Pawelczyk et al. 2021). In this work, we go one step further and ask ourselves: what happens if recourse is provided and implemented repeatedly? What types of dynamics are introduced, and how do different counterfactual generators compare in this context?

Research on Algorithmic Recourse has also so far typically addressed the issue from the perspective of a single individual. Arguably though, most real-world applications that warrant AR involve potentially large groups of individuals typically competing for scarce resources. Our work demonstrates that in such scenarios, choices made by or for a single individual are likely to affect the broader collective of individuals in ways that many current approaches to AR fail to account for. More specifically, we argue that a strict focus on minimizing the private costs to individuals may be too narrow an objective.

Figure 3.1 illustrates this idea for a binary problem involving a linear classifier and the counterfactual generator proposed by Wachter, Mittelstadt, and Russell (2017): the implementation of AR for a subset of samples from the negative class (orange) immediately leads to a visible domain shift in the (blue) target class (b), which in turn triggers a model shift (c). As this game of implementing AR and updating the classifier is repeated, the decision boundary moves away from training samples that were originally in the target class (d). We refer to these types of dynamics as **endogenous** because they are induced by the implementation of recourse itself. The term **macrodynamics** is borrowed from the economics literature and used to describe processes involving whole groups or societies.

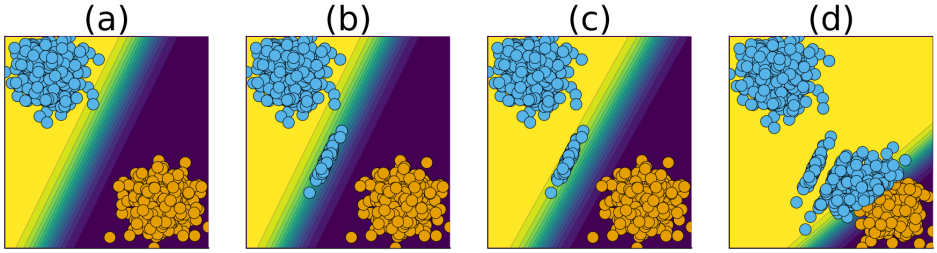


Figure 3.1. Dynamics in Algorithmic Recourse: (a) we have a simple linear classifier trained for binary classification where samples from the negative class ($y = 0$) are marked in orange and samples of the positive class ($y = 1$) are marked in blue; (b) the implementation of AR for a random subset of samples from the orange class leads to a noticeable domain shift, as the samples that have received recourse form a distinct new (blue) cluster; (c) as the classifier is retrained we observe a corresponding model shift; (d) as this process is repeated, the decision boundary moves away from the target class.

We think that these types of endogenous dynamics may be problematic and deserve our attention. From a purely technical perspective, we note the following: firstly, model shifts may inadvertently change classification outcomes for individuals who never received recourse. Secondly, we observe in Figure 3.1 that as the decision boundary moves in the direction of the non-target class, counterfactual paths become shorter. We think that in some practical applications, this can be expected to generate costs for involved stakeholders. To follow our argument, consider the

following two examples:

Example 3.1 (Consumer Credit)**.** Suppose Figure 3.1 relates to an automated decision-making system used by a retail bank to evaluate credit applicants with respect to their creditworthiness. Assume that the two features are meaningful in the sense that creditworthiness decreases in the bottom-right direction. Then we can think of the outcome in panel (d) as representing a situation where the bank supplies credit to more borrowers (blue), but these borrowers are on average less creditworthy and more of them can be expected to default on their loan. This represents a cost to the retail bank.

Example 3.2 (Student Admission)**.** Suppose Figure 3.1 relates to an automated decision-making system used by a university in its student admission process. Assume that the two features are meaningful in the sense that the likelihood of students completing their degree decreases in the bottom-right direction. Then we can think of the outcome in panel (b) as representing a situation where more students are admitted to university (blue), but they are more likely to fail their degree than students that were admitted in previous years. The university admission committee catches on to this and suspends its efforts to offer Algorithmic Recourse. This represents an opportunity cost to future student applicants, that may have derived utility from being offered recourse.

Both examples are exaggerated simplifications of potential real-world scenarios, but they serve to illustrate the point that recourse for one single individual may exert negative externalities on other individuals.

To the best of our knowledge, this is the first work investigating endogenous macrodynamics in AR. Our contributions to the state of knowledge are as follows: firstly, we posit a compelling argument that calls for a novel perspective on Algorithmic Recourse extending our focus from single individuals to groups (Sections Section 3.2 and Section 3.3). Secondly, we introduce an experimental framework extending previous work by Altmeyer, Deursen, and Liem (2023a) (Chapter 2), which enables us to study macrodynamics of Algorithmic Recourse through simulations that can be fully parallelized (Section Section 3.4). Thirdly, we use this framework to provide a first in-depth analysis of endogenous recourse dynamics induced by various popular counterfactual generators proposed in Wachter, Mittelstadt, and Russell (2017), Schut et al. (2021), Joshi et al. (2019), Mothilal, Sharma, and Tan (2020) and Antorán et al. (2020) (Section 3.5 and Section 3.6). Fourthly, given that we find a substantial impact of recourse, we propose and assess various mitigation strategies (Section Section 3.7). Finally, we discuss our findings in the broader context of the literature in Section Section 3.8, before pointing to some of the limitations of our work as well as avenues for future research in Section Section 3.9. Section Section 3.10 concludes.

3.2. BACKGROUND

In this section, we provide a review of the relevant literature. First, Subsection Section 3.2.1 discusses the existing research within the domain of Counterfactual Explanations and Algorithmic Recourse. Then, Subsection Section 3.2.2 presents some of the previous work on the measurement of data and model shifts.

3.2.1. ALGORITHMIC RECOURSE

A framework for Counterfactual Explanations was first proposed by Wachter, Mittelstadt, and Russell (2017) and has served as the baseline for many methodologies that have been proposed since then. Let $M : \mathcal{X} \mapsto \mathcal{Y}$ denote some pre-trained model that maps from inputs $X \in \mathcal{X}$ to outputs $Y \in \mathcal{Y}$. Then we are interested in minimizing the cost[1] $C = \text{cost}(x')$ incurred by individual x when moving to a counterfactual state x' such that the predicted outcome $M(x')$ corresponds to some target outcome y^*:

$$\min_{x' \in \mathcal{X}} \text{cost}(x') \quad \text{s. t.} \quad M(x') = y^* \tag{3.1}$$

For implementation purposes, [eq:obj] is typically approximated through regularization:

$$x' = \arg\min_{x'} \text{yloss}(M(x'), y^*) + \lambda \text{cost}(x') \tag{3.2}$$

In the baseline work (Wachter, Mittelstadt, and Russell 2017), the cost function is proxied by some distance metric based on the simple intuition that perturbations of x are costly to the individual. For models that are differentiable and produce smooth predictions, [eq:solution] can be solved through gradient descent. This summarizes the approach followed in Wachter, Mittelstadt, and Russell (2017) which we refer to simply as **Wachter**, the name of the first author, in the remainder of this paper.

Many approaches for the generation of Algorithmic Recourse have been described in the literature since 2017. An October 2020 survey laid out 60 algorithms that have been proposed since 2014 (Karimi et al. 2021). Another survey published around the same time described 29 algorithms (Verma et al. 2022). Different approaches vary primarily in terms of the objective functions they impose, how they optimize said objective (from brute force through gradient-based approaches to graph traversal algorithms), and how they ensure that certain requirements for CE are met. Regarding the latter, the literature has produced an extensive list of desiderata each addressing different needs. To name but a few, we are interested in generating counterfactuals that are close (Wachter, Mittelstadt, and Russell 2017), actionable (Ustun, Spangher, and Liu 2019), realistic (Schut et al. 2021), sparse,

[1] Equivalently, others have referred to this quantity as *complexity* or simply *distance*.

diverse (Mothilal, Sharma, and Tan 2020) and if possible causally founded (Karimi, Schölkopf, and Valera 2021).

Efforts so far have largely been directed at improving the quality of Counterfactual Explanations within a static context: given some pre-trained classifier $M : \mathcal{X} \mapsto \mathcal{Y}$, we are interested in generating one or multiple meaningful Counterfactual Explanations for some individual characterized by x. The ability of Counterfactual Explanations to handle dynamics like data and model shifts remains a largely unexplored research challenge at this point (Verma et al. 2022). We have been able to identify only one recent work that considers the implications of **exogenous** domain and model shifts in the context of AR (Upadhyay, Joshi, and Lakkaraju 2021). Exogenous shifts are strictly of external origin. For example, they might stem from data correction, temporal shifts or geospatial changes. Upadhyay, Joshi, and Lakkaraju (2021) propose ROAR: a framework for Algorithmic Recourse that evidently improves robustness to such exogenous shifts.

As mentioned earlier, research has so far also generally focused on generating counterfactuals for single individuals or instances. We have been able to identify only one existing work that investigates black-box model behavior towards a group of individuals (Carrizosa, Ramırez-Ayerbe, and Romero 2021). The authors propose an optimization framework that generates collective counterfactuals. We provide a motivation for doing so from the perspective of endogenous macrodynamics of Algorithmic Recourse.

3.2.2. DOMAIN AND MODEL SHIFTS

Much attention has been paid to the detection of dataset shifts – situations where the distribution of data changes over time. Rabanser, Günnemann, and Lipton (2019) suggest a framework to detect data drift from a minimal number of samples through the application of two-sample tests. This task is a generalization of the anomaly detection problem for large datasets, which aims to answer the question if two sets of samples could have been generated from the same probability distribution. Numerous approaches to anomaly detection have been summarized (Chandola, Banerjee, and Kumar 2009). Another well-established research topic is concept drift: situations where external variables influence the patterns between the input and the output of a model (Widmer and Kubat 1996). For instance, Gama et al. (2014) offer a review of the adaptive learning techniques which can handle concept drift. Less previous work is available on the related topic of model drift: changes in model performance over time. Nelson et al. (2015) review how resistant different machine learning models are to model drift. Ackerman et al. (2021) offer a method to detect changes in model performance when ground truth is not available.

In the context of Algorithmic Recourse, domain and model shifts were first brought up by the authors behind ROAR (Upadhyay, Joshi, and Lakkaraju 2021). In their work, they refer to model shifts as simply any perturbation Δ to the parameters of the model in question: M. While this also sets the baseline for our analysis here, it

is worth noting that in Upadhyay, Joshi, and Lakkaraju (2021) these perturbations are mechanically introduced. In contrast, we are interested in quantifying model shifts that arise endogenously as part of a dynamic recourse process. In addition to quantifying the magnitude of shifts Δ, we aim to also analyze the characteristics of changes to the model, such as the position of the decision boundary and the overall decisiveness of the model. We have not been able to identify previous work on this topic.

3.2.3. BENCHMARKING COUNTERFACTUAL GENERATORS

Despite the large and growing number of approaches to counterfactual search, there have been surprisingly few benchmark studies that compare different methodologies. This may be partially due to limited software availability in this space. Recent work has started to address this gap: firstly, de Oliveira and Martens (2021) run a large benchmarking study using different algorithmic approaches and numerous tabular datasets; secondly, Pawelczyk et al. (2021) introduce a Python framework—CARLA—that can be used to apply and benchmark different methodologies; finally, `CounterfactualExplanations.jl` (Chapter 2) provides an extensible and fast implementation in Julia. Since the experiments presented here involve extensive simulations, we have relied on and extended the Julia implementation due to the associated performance benefits. In particular, we have built a framework on top of `CounterfactualExplanations.jl` that extends the functionality from static benchmarks to simulation experiments: `AlgorithmicRecourseDynamics.jl`[2]. The core concepts implemented in that package reflect what is presented in Section Section 3.4 of this paper.

3.3. GRADIENT-BASED RECOURSE REVISITED

In this section, we first set out a generalized framework for gradient-based counterfactual search that encapsulates the various Individual Recourse methods we have chosen to use in our experiments (Section Section 3.3.1). We then introduce the notion of a hidden external cost in Algorithmic Recourse and extend the existing framework to explicitly address this cost in the counterfactual search objective (Section Section 3.3.2).

3.3.1. FROM INDIVIDUAL RECOURSE ...

We have chosen to focus on gradient-based counterfactual search for two reasons: firstly, they can be seen as direct descendants of our baseline method (Wachter); secondly, gradient-based search is particularly well-suited for differentiable black-box models like deep neural networks, which we focus on in this work. In particular,

[2]The code has been released as a package: https://github.com/pat-alt/AlgorithmicRecourseDynamics.jl.

we include the following generators in our simulation experiments below: **REVISE** (Joshi et al. 2019), **CLUE** (Antorán et al. 2020), **DiCE** (Mothilal, Sharma, and Tan 2020) and a greedy approach that relies on probabilistic models (Schut et al. 2021). Our motivation for including these different generators in our analysis is that they all offer slightly different approaches to generating meaningful counterfactuals for differentiable black-box models. We hypothesize that generating more **meaningful** counterfactuals should mitigate the endogenous dynamics illustrated in Figure 3.1 in Section 3.1. This intuition stems from the underlying idea that more meaningful counterfactuals are generated by the same or at least a very similar data-generating process as the observed data. All else equal, counterfactuals that fulfil this basic requirement should be less prone to trigger shifts.

As we will see next, all of them can be described by the following generalized form of Equation 3.2:

$$\mathbf{s}' = \arg\min_{\mathbf{s}' \in \mathcal{S}} \{\text{yloss}(M(f(\mathbf{s}')), y^*) + \lambda \text{cost}(f(\mathbf{s}'))\} \tag{3.3}$$

Here $\mathbf{s}' = \{s'_k\}_K$ is a K-dimensional array of counterfactual states and $f : \mathcal{S} \mapsto \mathcal{X}$ maps from the counterfactual state space to the feature space. In Wachter, the state space is the feature space: f is the identity function and the number of counterfactuals K is one. Both REVISE and CLUE search counterfactuals in some latent space $\mathcal{S}$ instead of the feature space directly. The latent embedding is learned by a separate generative model that is tasked with learning the data-generating process (DGP) of X. In this case, f in Equation 5.1 corresponds to the decoder part of the generative model, that is the function that maps back from the latent space to inputs. Provided the generative model is well-specified, traversing the latent embedding typically yields meaningful counterfactuals since they are implicitly generated by the (learned) DGP (Joshi et al. 2019).

CLUE distinguishes itself from REVISE and other counterfactual generators in that it aims to minimize the predictive uncertainty of the model in question, M. To quantify predictive uncertainty, Antorán et al. (2020) rely on entropy estimates for probabilistic models. The greedy approach proposed by Schut et al. (2021), which we refer to as **Greedy**, also works with the subclass of models $\tilde{\mathcal{M}} \subset \mathcal{M}$ that can produce predictive uncertainty estimates. The authors show that in this setting the cost function $\text{cost}(\cdot)$ in Equation 5.1 is redundant and meaningful counterfactuals can be generated in a fast and efficient manner through a modified Jacobian-based Saliency Map Attack (JSMA). Schut et al. (2021) also show that by maximizing the predicted probability of x' being assigned to target class y^*, we also implicitly minimize predictive entropy (as in CLUE). In that sense, CLUE can be seen as equivalent to REVISE in the Bayesian context and we shall therefore refer to both approaches collectively as **Latent Space** generators[3].

[3] In fact, there are several other recently proposed approaches to counterfactual search that also broadly fall in this same category. They largely differ with respect to the chosen generative model: for example, the generator proposed by Dombrowski, Gerken, and Kessel (2021) relies on normalizing flows.

Finally, DiCE (Mothilal, Sharma, and Tan 2020) distinguishes itself from all other generators considered here in that it aims to generate a diverse set of $K > 1$ counterfactuals. Wachter, Mittelstadt, and Russell (2017) show that diverse outcomes can in principle be achieved simply by rerunning counterfactual search multiple times using stochastic gradient descent (or by randomly initializing the counterfactual)[4]. In Mothilal, Sharma, and Tan (2020) diversity is explicitly proxied via Determinantal Point Processes (DDP): the authors introduce DDP as a component of the cost function $\text{cost}(\mathbf{s}')$ and thereby produce counterfactuals $s_1, ..., s_K$ that look as different from each other as possible. The implementation of DiCE in our library of choice—`CounterfactualExplanations.jl`—uses that exact approach. It is worth noting that for $k = 1$, DiCE reduces to Wachter since the DDP is constant and therefore does not affect the objective function in Equation 5.1.

3.3.2. ... TOWARDS COLLECTIVE RECOURSE

All of the different approaches introduced above tackle the problem of Algorithmic Recourse from the perspective of one single individual[5]. To explicitly address the issue that Individual Recourse may affect the outcome and prospect of other individuals, we propose to extend Equation 5.1 as follows:

$$\begin{aligned} \mathbf{s}' = \arg\min_{\mathbf{s}' \in \mathcal{S}} \{ & \text{yloss}(M(f(\mathbf{s}')), y^*) \\ & + \lambda_1 \text{cost}(f(\mathbf{s}')) + \lambda_2 \text{extcost}(f(\mathbf{s}')) \} \end{aligned} \tag{3.4}$$

Here $\text{cost}(f(\mathbf{s}'))$ denotes the proxy for private costs faced by the individual as before and λ_1 governs to what extent that private cost ought to be penalized. The newly introduced term $\text{extcost}(f(\mathbf{s}'))$ is meant to capture and address external costs incurred by the collective of individuals in response to changes in $\mathbf{s}'$. The underlying concept of private and external costs is borrowed from Economics and well-established in that field: when the decisions or actions by some individual market participant generate external costs, then the market is said to suffer from negative externalities and is considered inefficient (Pindyck and Rubinfeld 2014). We think that this concept describes the endogenous dynamics of algorithmic recourse observed here very well. As with Individual Recourse, the exact choice of $\text{extcost}(\cdot)$ is not obvious, nor do we intend to provide a definitive answer in this work, if such even exists. That being said, we do propose a few potential mitigation strategies in Section 3.7.

[4] Note that Equation 5.1 naturally lends itself to that idea: setting K to some value greater than one and using the Wachter objective essentially boils down to computing multiple counterfactuals in parallel. Here, $yloss(\cdot)$ is first broadcasted over elements of $\mathbf{s}'$ and then aggregated. This is exactly how counterfactual search is implemented in `CounterfactualExplanations.jl`.

[5] DiCE recognizes that different individuals may have different objective functions, but it does not address the interdependencies between different individuals.

3.4. MODELLING ENDOGENOUS MACRODYNAMICS IN AR

In the following, we describe the framework we propose for modelling and analyzing endogenous macrodynamics in Algorithmic Recourse. We introduce this framework with the ambition to shed light on the following research questions:

Research Question 3.1 (Endogenous Shifts)**.** Does the repeated implementation of recourse provided by state-of-the-art generators lead to shifts in the domain and model?

Research Question 3.2 (Costs)**.** If so, are these dynamics substantial enough to be considered costly to stakeholders involved in real-world automated decision-making processes?

Research Question 3.3 (Heterogeneity)**.** Do different counterfactual generators yield significantly different outcomes in this context? Furthermore, is there any heterogeneity concerning the chosen classifier and dataset?

Research Question 3.4 (Drivers)**.** What are the drivers of endogenous dynamics in Algorithmic Recourse?

Below we first describe the basic simulations that were generated to produce the findings in this work and also constitute the core of `AlgorithmicRecourseDynamics.jl`—the Julia package we introduced earlier. The remainder of this section then introduces various evaluation metrics that can be used to benchmark different counterfactual generators with respect to how they perform in the dynamic setting.

3.4.1. SIMULATIONS

The dynamics illustrated in Figure 3.1 were generated through a simple experiment that aims to simulate the process of Algorithmic Recourse in practice. We begin in the static setting at time $t = 0$: firstly, we have some binary classifier M that was pre-trained on data $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, where $\mathcal{D}_0$ and $\mathcal{D}_1$ denote samples in the non-target and target class, respectively; secondly, we generate recourse for a random batch of B individuals in the non-target class ($\mathcal{D}_0$). Note that we focus our attention on classification problems since classification poses the most common use-case for recourse[6].

In order to simulate the dynamic process, we suppose that the model M is retrained following the actual implementation of recourse in time $t = 0$. Following the update to the model, we assume that at time $t = 1$ recourse is generated for yet another random subset of individuals in the non-target class. This process is repeated for

[6]To keep notation simple, we have also restricted ourselves to binary classification here, but `AlgorithmicRecourseDynamics.jl` can also be used for multi-class problems.

a number of time periods T. To get a clean read on endogenous dynamics we keep the total population of samples closed: we allow existing samples to move from factual to counterfactual states but do not allow any entirely new samples to enter the population. The experimental setup is summarized in Algorithm 3.1.

Algorithm 3.1 Simulation Experiment.

1: **procedure** Experiment$(M, \mathcal{D}, G)$
2: $\quad E \leftarrow \emptyset$ ▷ Initialize evaluation E.
3: $\quad t \leftarrow 0$
4: $\quad$ **while** $t < T$ **do**
5: $\quad\quad$ batch $\subset \mathcal{D}_0$ ▷ Sample from $\mathcal{D}_0$ (assignment).
6: $\quad\quad$ batch $\leftarrow G(\text{batch})$ ▷ Generate counterfactuals.
7: $\quad\quad M \leftarrow M(\mathcal{D})$ ▷ Retrain model.
8: $\quad\quad E \leftarrow \text{eval}(M, \mathcal{D}) \cup E$ ▷ Update evaluation.
9: $\quad\quad t \leftarrow t + 1$ ▷ Increment t.
10: $\quad$ **end while**
11: $\quad$ **return** $E, M, \mathcal{D}$
12: **end procedure**

Note that the operation in line 4 is an assignment, rather than a copy operation, so any updates to 'batch' will also affect $\mathcal{D}$. The function eval$(M, \mathcal{D})$ loosely denotes the computation of various evaluation metrics introduced below. In practice, these metrics can also be computed at regular intervals as opposed to every round.

Along with any other fixed parameters affecting the counterfactual search, the parameters T and B are assumed as given in Algorithm 3.1. Still, it is worth noting that the higher these values, the more factual instances undergo recourse throughout the entire experiment. Of course, this is likely to lead to more pronounced domain and model shifts by time T. In our experiments, we choose the values such that the majority of the negative instances from the initial dataset receive recourse. As we compute evaluation metrics at regular intervals throughout the procedure, we can also verify the impact of recourse when it is implemented for a smaller number of individuals.

Algorithm 3.1 summarizes the proposed simulation experiment for a given dataset $\mathcal{D}$, model M and generator G, but naturally, we are interested in comparing simulation outcomes for different sources of data, models and generators. The framework we have built facilitates this, making use of multithreading in order to speed up computations. Holding the initial model and dataset constant, the experiments are run for all generators, since our primary concern is to benchmark different recourse methods. To ensure that each generator is faced with the same initial conditions in each round t, the candidate batch of individuals from the non-target class is randomly drawn from the intersection of all non-target class individuals across all experiments $\{\text{Experiment}(M, \mathcal{D}, G)\}_{j=1}^{J}$ where J is the total number of generators.

3.4.2. EVALUATION METRICS

We formulate two desiderata for the set of metrics used to measure domain and model shifts induced by recourse. First, the metrics should be applicable regardless of the dataset or classification technique so that they allow for the meaningful comparison of the generators in various scenarios. As knowledge of the underlying probability distribution is rarely available, the metrics should be empirical and non-parametric. This further ensures that we can also measure large datasets by sampling from the available data. Moreover, while our study was conducted in a two-class classification setting, our choice of metrics should remain applicable in future research on multi-class recourse problems. Second, the set of metrics should allow capturing various aspects of the previously mentioned magnitude, path, and pace of changes while remaining as small as possible.

3.4.2.1. DOMAIN SHIFTS

To quantify the magnitude of domain shifts we rely on an unbiased estimate of the squared population **Maximum Mean Discrepancy (MMD)** given as:

$$\begin{aligned} MMD(X', \tilde{X}') = & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) \\ & + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(\tilde{x}_i, \tilde{x}_j) \\ & - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, \tilde{x}_j) \end{aligned} \tag{3.5}$$

where $X = \{x_1, ..., x_m\}$, $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_n\}$ represent independent and identically distributed samples drawn from probability distributions $\mathcal{X}$ and $\tilde{\mathcal{X}}$ respectively (Gretton et al. 2012). MMD is a measure of the distance between the kernel mean embeddings of $\mathcal{X}$ and $\tilde{\mathcal{X}}$ in a Reproducing Kernel Hilbert Space, $\mathcal{H}$ (Berlinet and Thomas-Agnan 2011). An important consideration is the choice of the kernel function $k(\cdot, \cdot)$. In our implementation, we make use of a Gaussian kernel with a constant length-scale parameter of 0.5. As the Gaussian kernel captures all moments of distributions $\mathcal{X}$ and $\tilde{\mathcal{X}}$, we have that $MMD(X, \tilde{X}) = 0$ if and only if $X = \tilde{X}$. Conversely, larger values $MMD(X, \tilde{X}) > 0$ indicate that it is more likely that $\mathcal{X}$ and $\tilde{\mathcal{X}}$ are different distributions. In our context, large values, therefore, indicate that a domain shift indeed seems to have occurred.

To assess the statistical significance of the observed shifts under the null hypothesis that samples X and $\tilde{X}$ were drawn from the same probability distribution, we follow Arcones and Gine (1992). To that end, we combine the two samples and generate a large number of permutations of $X + \tilde{X}$. Then, we split the permuted data into

two new samples X' and $\tilde{X}'$ having the same size as the original samples. Under the null hypothesis, we should have that $MMD(X', \tilde{X}')$ be approximately equal to $MMD(X, \tilde{X})$. The corresponding p-value can then be calculated by counting how often these two quantities are not equal.

We calculate the MMD for both classes individually based on the ground truth labels, i.e. the labels that samples were assigned in time $t = 0$. Throughout our experiments, we generally do not expect the distribution of the negative class to change over time – application of recourse reduces the size of this class, but since individuals are sampled uniformly the distribution should remain unaffected. Conversely, unless a recourse generator can perfectly replicate the original probability distribution, we expect the MMD of the positive class to increase. Thus, when discussing MMD, we generally mean the shift in the distribution of the positive class.

3.4.2.2. MODEL SHIFTS

As our baseline for quantifying model shifts, we measure perturbations to the model parameters at each point in time t following Upadhyay, Joshi, and Lakkaraju (2021). We define $\Delta = ||\theta_{t+1} - \theta_t||^2$, that is the euclidean distance between the vectors of parameters before and after retraining the model M. We shall refer to this baseline metric simply as **Perturbations**.

Furthermore, we extend the metric in Equation 5.8 to quantify model shifts. Specifically, we introduce **Predicted Probability MMD (PP MMD)**: instead of applying Equation 5.8 to features directly, we apply it to the predicted probabilities assigned to a set of samples by the model M. If the model shifts, the probabilities assigned to each sample will change; again, this metric will equal 0 only if the two classifiers are the same. We compute PP MMD in two ways: firstly, we compute it over samples drawn uniformly from the dataset, and, secondly, we compute it over points spanning a mesh grid over a subspace of the entire feature space. For the latter approach, we bound the subspace by the extrema of each feature. While this approach is theoretically more robust, unfortunately, it suffers from the curse of dimensionality, since it becomes increasingly difficult to select enough points to overcome noise as the dimension D grows.

As an alternative to PP MMD, we use a pseudo-distance for the **Disagreement Coefficient** (Disagreement). This metric was introduced in Hanneke (2007) and estimates $p(M(x) \neq M'(x))$, that is the probability that two classifiers disagree on the predicted outcome for a randomly chosen sample. Thus, it is not relevant whether the classification is correct according to the ground truth, but only whether the sample lies on the same side of the two respective decision boundaries. In our context, this metric quantifies the overlap between the initial model (trained before the application of AR) and the updated model. A Disagreement Coefficient unequal to zero is indicative of a model shift. The opposite is not true: even if the Disagreement Coefficient is equal to zero, a model shift may still have occurred. This is one reason why PP MMD is our preferred metric.

We further introduce **Decisiveness** as a metric that quantifies the likelihood that a model assigns a high probability to its classification of any given sample. We define the metric simply as $\frac{1}{N}\sum_{i=0}^{N}(\sigma(M(x)) - 0.5)^2$ where $M(x)$ are predicted logits from a binary classifier and σ denotes the sigmoid function. This metric provides an unbiased estimate of the binary classifier's tendency to produce high-confidence predictions in either one of the two classes. Although the exact values for this metric are not important for our study, they can be used to detect model shifts. If decisiveness changes over time, then this is indicative of the decision boundary moving towards either one of the two classes. A potential caveat of this metric in the context of our experiments is that it will to some degree get inflated simply through retraining the model.

Finally, we also take a look at the out-of-sample **Performance** of our models. To this end, we compute their F-score on a test sample that we leave untouched throughout the experiment.

3.5. EXPERIMENT SETUP

This section presents the exact ingredients and parameter choices describing the simulation experiments we ran to produce the findings presented in the next section (Section 3.6). For convenience, we use Algorithm 3.1 as a template to guide us through this section. A few high-level details upfront: each experiment is run for a total of $T = 50$ rounds, where in each round we provide recourse to five per cent of all individuals in the non-target class, so $B_t = 0.05 * N_t^{\mathcal{D}_0}$. All classifiers and generative models are retrained for 10 epochs in each round t of the experiment. Rather than retraining models from scratch, we initialize all parameters at their previous levels $(t-1)$ and backpropagate for 10 epochs using the new training data as inputs into the existing model. Evaluation metrics are computed and stored every 10 rounds. To account for noise, each individual experiment is repeated five times[7].

3.5.1. M—CLASSIFIERS AND GENERATIVE MODELS

For each dataset and generator, we look at three different types of classifiers, all of them built and trained using `Flux.jl` (Michael Innes et al. 2018): firstly, a simple linear classifier—**Logistic Regression**—implemented as a single linear layer with sigmoid activation; secondly, a multilayer perceptron (**MLP**); and finally, a **Deep Ensemble** composed of five MLPs following Lakshminarayanan, Pritzel, and Blundell (2017) that serves as our only probabilistic classifier. We have chosen to work with deep ensembles both for their simplicity and effectiveness at modelling predictive uncertainty. They are also the model of choice in Schut et al. (2021).

[7] In the current implementation, we use the same train-test split each time to only account for stochasticity associated with randomly selecting individuals for recourse. An interesting alternative may be to also perform data splitting each time, thereby adding a layer of randomness.

The network architectures are kept simple (top half of Table 3.1), since we are only marginally concerned with achieving good initial classifier performance.

Table 3.1. Neural network architectures and training parameters.

	Data	Hidden Dim.	Latent Dim.	Hidden Layers	Batch	Dropout	Epochs
MLP							
	Synthetic	32	-	1	-	-	100
	Real-World	64	-	2	500	0.1	100
VAE							
	Synthetic	32	2	1	-	-	100
	Real-World	32	8	1	-	-	250

The Latent Space generator relies on a separate generative model. Following the authors of both REVISE and CLUE we use Variational Autoencoders (**VAE**) for this purpose. As with the classifiers, we deliberately choose to work with fairly simple architectures (bottom half of Table 3.1). More expressive generative models generally also lead to more meaningful counterfactuals produced by Latent Space generators. But in our view, this should simply be considered as a vulnerability of counterfactual generators that rely on surrogate models to learn realistic representations of the underlying data.

3.5.2. $\mathcal{D}$—DATA

We have chosen to work with both synthetic and real-world datasets. Using synthetic data allows us to impose distributional properties that may affect the resulting recourse dynamics. Following Upadhyay, Joshi, and Lakkaraju (2021), we generate synthetic data in $\mathbb{R}^2$ to also allow for a visual interpretation of the results. Real-world data is used in order to assess if endogenous dynamics also occur in higher-dimensional settings.

3.5.2.1. SYNTHETIC DATA

We use four synthetic binary classification datasets consisting of 1000 samples each: **Overlapping**, **Linearly Separable**, **Circles** and **Moons** (Figure 3.2).

Ex-ante we expect to see that by construction, Wachter will create a new cluster of counterfactual instances in the proximity of the initial decision boundary as we saw in Figure 3.1. Thus, the choice of a black-box model may have an impact on the counterfactual paths. For generators that use latent space search (REVISE (Joshi et al. 2019), CLUE (Antorán et al. 2020)) or rely on (and have access to) probabilistic models (CLUE (Antorán et al. 2020), Greedy (Schut et al. 2021)) we expect that

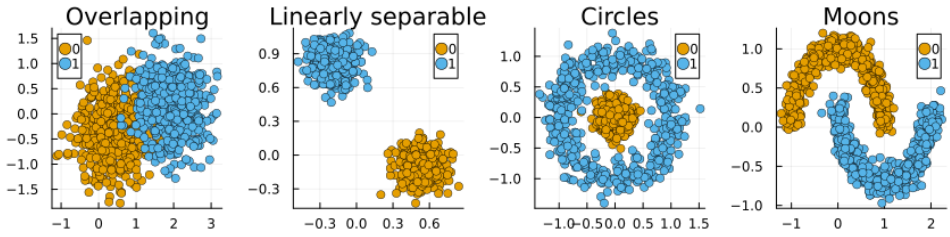


Figure 3.2. Synthetic classification datasets used in our experiments. Samples from the negative class ($y = 0$) are marked in blue while samples of the positive class ($y = 1$) are marked in orange.

counterfactuals will end up in regions of the target domain that are densely populated by training samples. Of course, this expectation hinges on how effective said probabilistic models are at capturing predictive uncertainty. Finally, we expect to see the counterfactuals generated by DiCE to be diversely spread around the feature space inside the target class[8]. In summary, we expect that the endogenous shifts induced by Wachter outsize those of all other generators since Wachter is not explicitly concerned with generating what we have defined as meaningful counterfactuals.

3.5.3. REAL-WORLD DATA

We use three different real-world datasets from the Finance and Economics domain, all of which are tabular and can be used for binary classification. Firstly, we use the **Give Me Some Credit** dataset which was open-sourced on Kaggle for the task to predict whether a borrower is likely to experience financial difficulties in the next two years (Kaggle 2011), originally consisting of 250,000 instances with 11 numerical attributes. Secondly, we use the **UCI defaultCredit** dataset (Yeh and Lien 2009), a benchmark dataset that can be used to train binary classifiers to predict the binary outcome variable of whether credit card clients default on their payment. In its raw form, it consists of 23 explanatory variables: 4 categorical features relating to demographic attributes and 19 continuous features largely relating to individuals' payment histories and amount of credit outstanding. Both datasets have been used in the literature on AR before (see for example Pawelczyk et al. (2021), Joshi et al. (2019) and Ustun, Spangher, and Liu (2019)), presumably because they constitute real-world classification tasks involving individuals that compete for access to credit.

As a third dataset, we include the **California Housing** dataset derived from the 1990 U.S. census and sourced through scikit-learn (Pace and Barry 1997). It consists of 8 continuous features that can be used to predict the median house price for California

[8]As we mentioned earlier, the diversity constraint used by DiCE is only effective when at least two counterfactuals are being generated. We have therefore decided to always generate 5 counterfactuals for each generator and randomly pick one of them.

districts. The continuous outcome variable is binarized as $\tilde{y} = \mathbb{1}_{y>\text{median}(Y)}$ indicating whether the median house price of a given district is above the median of all districts. While we have not seen this dataset used in the previous literature on AR, others have used the Boston Housing dataset in a similar fashion (Schut et al. 2021). We initially also conducted experiments on that dataset, but eventually discarded it due to surrounding ethical concerns (Carlisle 2019).

Since the simulations involve generating counterfactuals for a significant proportion of the entire sample of individuals, we have randomly undersampled each dataset to yield balanced subsamples consisting of 5,000 individuals each. We have also standardized all continuous explanatory features since our chosen classifiers are sensitive to scale.

3.5.4. G—GENERATORS

All generators introduced earlier are included in the experiments: Wachter (Wachter, Mittelstadt, and Russell 2017), REVISE (Joshi et al. 2019), CLUE (Antorán et al. 2020), DiCE (Mothilal, Sharma, and Tan 2020) and Greedy (Schut et al. 2021). In addition, we introduce two new generators in Section Section 3.7 that directly address the issue of endogenous domain and model shifts. We also test to what extent it may be beneficial to combine ideas underlying the various generators.

3.6. EXPERIMENTS

Below, we first present our main experimental findings regarding these questions. We conclude this section with a brief recap providing answers to all of these questions.

3.6.1. ENDOGENOUS MACRODYNAMICS

We start this section off with the key high-level observations. Across all datasets (synthetic and real), classifiers and counterfactual generators we observe either most or all of the following dynamics at varying degrees:

- Statistically significant domain and model shifts as measured by MMD.
- A deterioration in out-of-sample model performance as measured by the F-Score evaluated on a test sample. In many cases this drop in performance is substantial.
- Significant perturbations to the model parameters as well as an increase in the model's decisiveness.
- Disagreement between the original and retrained model, in some cases large.

There is also some clear heterogeneity across the results:

- The observed dynamics are generally of the highest magnitude for the linear classifier. Differences in results for the MLP and Deep Ensemble are mostly negligible.
- The reduction in model performance appears to be most severe when classes are not perfectly separable or the initial model performance was weak, to begin with.
- Except for the Greedy generator, all other generators generally perform somewhat better overall than the baseline (Wachter) as expected.

Focusing first on synthetic data, Figure 3.3 presents our findings for the dataset with overlapping classes. It shows the resulting values for some of our evaluation metrics at the end of the experiment, after all $T = 50$ rounds, along with error bars indicating the variation across folds.

The top row shows the estimated domain shifts. While it is difficult to interpret the exact magnitude of MMD, we can see that the values are different from zero and there is essentially no variation across our five folds. For the domain shifts, the Greedy generator induces the smallest shifts, possibly because it only ever affects one feature per iteration. In general, we have observed the opposite[9].

The second row shows the estimated model shifts, where here we have used the grid approach explained earlier. As with the domain shifts, the observed values are clearly different from zero and variation across folds is once again small. In this case, the results for this particular dataset very much reflect the broader patterns we have observed: Latent Space (LS) generators induce the smallest shifts, while DiCE and Greedy are generally on par with generic search (Wachter)[10].

The same broad pattern also emerges in the third row: we observe the smallest deterioration in model performance for LS generators with a reduction in the F-Score of at most 2 percentage points. For DiCE and Wachter, the reduction in performance is up to 5 percentage points for non-linear models and up to nearly 15 percentage points for the linear classifier[11]. Related to this, the third row indicates that the retrained classifiers disagree with their initial counterparts on the classification of up to nearly 20 per cent of the individuals[12]. We also note that the final classifiers are more decisive, although as we noted earlier this may to some extent just be a byproduct of retraining the model throughout the experiment.

Figure 3.3 also indicates that the estimated effects are strongest for the simplest linear classifier, a pattern that we have observed fairly consistently. Conversely,

[9] For the Linearly Separable data, Greedy induces much stronger shifts than other generators. For the Moons dataset, this also holds for non-linear models.

[10] In the original article, we mistakenly stated that Greedy generally introduces the most substantial shifts. What we should have stated instead is that the results for Greedy exert the highest levels of volatility across datasets and metrics, where in some cases Greedy produces the worst results out of all generators, while in other cases it induces the smalles shifts.

[11] We have provided some more detail here than in the original article.

[12] The original article incorrectly stated “nearly 25 percent”.

there is virtually no difference in outcomes between the deep ensemble and the MLP. It is possible that the deep ensembles simply fail to capture predictive uncertainty well and hence counterfactual generators like Greedy, which explicitly addresses this quantity, fail to work as expected.

The findings for the other synthetic datasets are broadly consistent with the observations above (see appendix). For the Moons data, the same broad patterns emerge, although in this case, the Greedy generator induces comparably strong shifts in some cases. For the Circles data, model shifts and performance deterioration are quantitatively much smaller than what we can observe in Figure 3.3 and in many cases insignificant. For the Linearly Separable data we also find substantial domain and model shifts[13].

Finally, it is also worth noting that the observed dynamics and patterns are consistent throughout the experiment. That is to say that we start observing shifts already after just a few rounds and these tend to increase proportionately for the different generators over the course of the experiment.

Turning to the real-world data we will go through the findings presented in Figure 3.4, where each column corresponds to one of the three data sets. The results shown here are for the deep ensemble, which once again largely resemble those for the MLP. Starting from the top row, we find domain shifts of varying magnitudes that are in many cases statistically significant (see appendix for details). Latent Space search induces shifts that are orders of magnitude higher than for the other generators, which generally induce significant but small shifts.

Model shifts are shown in the middle row of Figure 3.4: the estimated PP MMD is statistically significant across the board and in some cases much larger than in others. We find no evidence that LS search helps to mitigate model shifts, as we did before for the synthetic data. Since these real-world datasets are arguably more complex than the synthetic data, the generative model can be expected to have a harder time learning the data-generating process and hence this increased difficulty appears to affect the performance of REVISE/CLUE.

The out-of-sample model performance also deteriorates across the board and substantially so: the largest average reduction in F-Scores of more than 10 percentage points is observed for the Credit Default dataset. For this dataset we achieved the lowest initial model performance, indicating once again that weaker classifiers may be more exposed to endogenous dynamics. As with the synthetic data, the estimates for logistic regression are qualitatively in line with the above, but quantitatively even more pronounced.

To recap, we answer our research questions: firstly, endogenous dynamics do emerge in our experiments (RQ 3.1) and we find them substantial enough to be considered costly (RQ 3.2); secondly, the choice of the counterfactual generator matters, with Latent Space search generally having a dampening effect (RQ 3.3). The observed

[13] In the original article we mistakenly stated that for the linearly separable data we observe "almost no reduction in model performance", which is only true for the non-linear models.

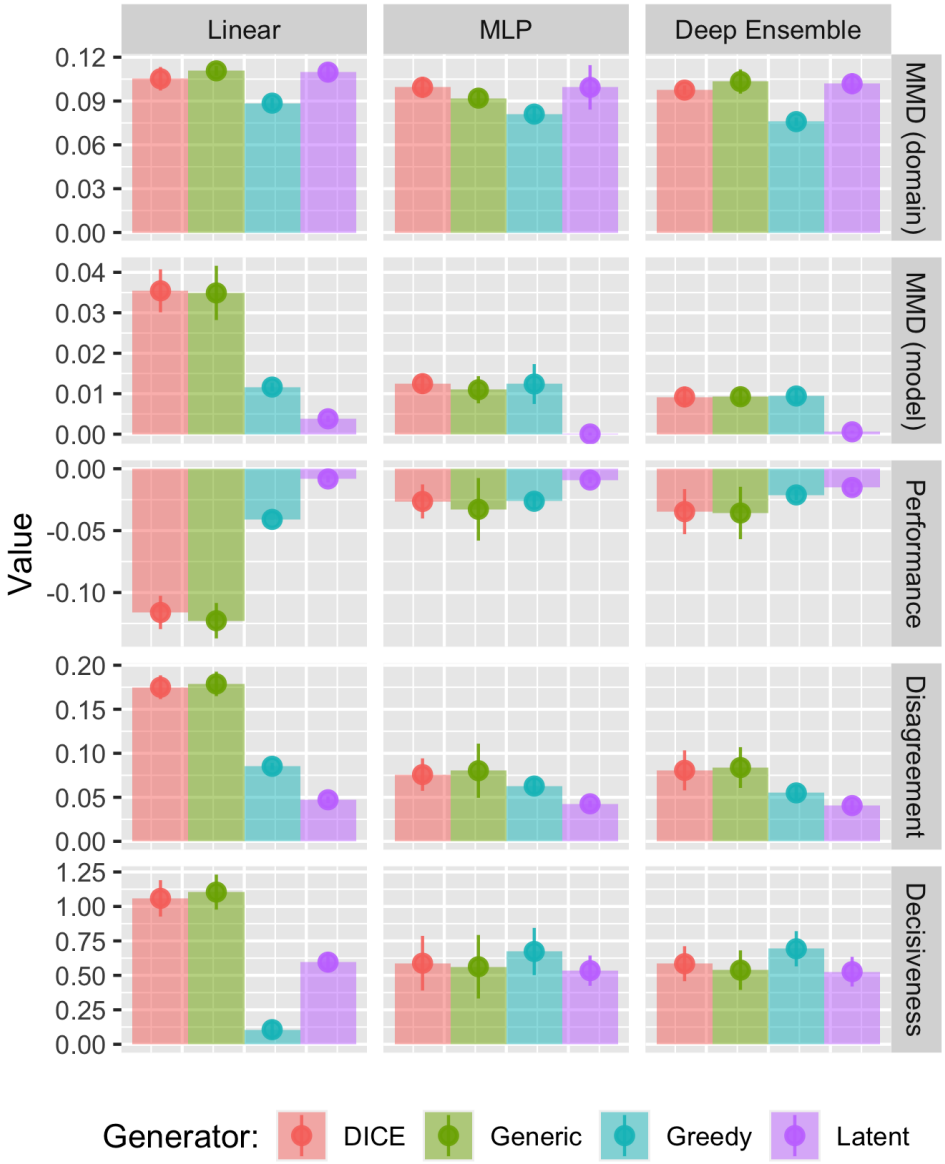


Figure 3.3. Results for synthetic data with overlapping classes. The shown model MMD (PP MMD) was computed over a mesh grid of 1,000 points. Error bars indicate the standard deviation across folds.

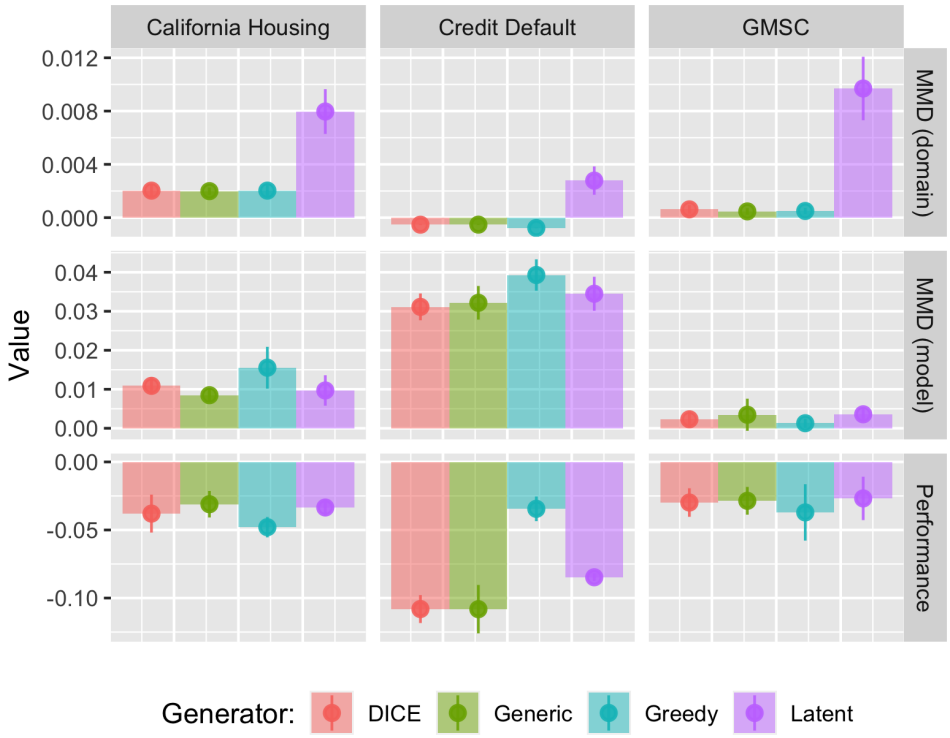


Figure 3.4. Results for deep ensemble using real-world datasets. The shown model MMD (PP MMD) was computed over actual samples, rather than a mesh grid. Error bars indicate the standard deviation across folds.

dynamics, therefore, seem to be driven by a discrepancy between counterfactual outcomes that minimize costs to individuals and outcomes that comply with the data-generating process (RQ 3.4).

3.7. MITIGATION STRATEGIES AND EXPERIMENTS

Having established in the previous section that endogenous macrodynamics in AR are substantial enough to warrant our attention, in this section we ask ourselves:

Research Question 3.5 (Mitigation Strategies)**.** What are potential mitigation strategies with respect to endogenous macrodynamics in AR?

We propose and test several simple mitigation strategies. All of them essentially boil down to one simple principle: to avoid domain and model shifts, the generated counterfactuals should comply as much as possible with the true data-generating process. This principle is really at the core of Latent Space (LS) generators, and hence it is not surprising that we have found these types of generators to perform comparably well in the previous section. But as we have mentioned earlier, generators that rely on separate generative models carry an additional computational burden and, perhaps more importantly, their performance hinges on the performance of said generative models. Fortunately, it turns out that we can use a number of other, much simpler strategies.

3.7.1. MORE CONSERVATIVE DECISION THRESHOLDS

The most obvious and trivial mitigation strategy is to simply choose a higher decision threshold γ. This threshold determines when a counterfactual should be considered valid. Under $\gamma = 0.5$, counterfactuals will end up near the decision boundary by construction. Since this is the region of maximal aleatoric uncertainty, the classifier is bound to be thrown off. By setting a more conservative threshold, we can avoid this issue to some extent. A drawback of this approach is that a classifier with high decisiveness may classify samples with high confidence even far away from the training data.

3.7.2. CLASSIFIER PRESERVING ROAR (CLAPROAR)

Another strategy draws inspiration from ROAR (Upadhyay, Joshi, and Lakkaraju 2021): to preserve the classifier, we propose to explicitly penalize the loss it incurs when evaluated on the counterfactual x' at given parameter values. Recall that extcost$(\cdot)$ denotes what we had defined as the external cost in Equation 3.4. Formally, we let

$$\text{extcost}(f(\mathbf{s}')) = l(M(f(\mathbf{s}')), y') \tag{3.6}$$

for each counterfactual k where l denotes the loss function used to train M. This approach, which we refer to as **ClaPROAR**, is based on the intuition that (endogenous) model shifts will be triggered by counterfactuals that increase classifier loss. It is closely linked to the idea of choosing a higher decision threshold, but is likely better at avoiding the potential pitfalls associated with highly decisive classifiers. It also makes the private vs. external cost trade-off more explicit and hence manageable.

3.7.3. Gravitational Counterfactual Explanations

Yet another strategy extends Wachter as follows: instead of only penalizing the distance of the individuals' counterfactual to its factual, we propose penalizing its distance to some sensible point in the target domain, for example, the subsample average $\bar{x}^* = \text{mean}(x)$, $x \in \mathcal{D}_1$:

$$\text{extcost}(f(\mathbf{s}')) = \text{dist}(f(\mathbf{s}'), \bar{x}^*) \tag{3.7}$$

Once again we can put this in the context of Equation 3.4: the former penalty can be thought of here as the private cost incurred by the individual, while the latter reflects the external cost incurred by other individuals. Higher choices of λ_2 relative to λ_1 will lead counterfactuals to gravitate towards the specified point $\bar{x}^*$ in the target domain. In the remainder of this paper, we will therefore refer to this approach as **Gravitational** generator, when we investigate its usefulness for mitigating endogenous macrodynamics[14].

Figure 3.5 shows an illustrative example that demonstrates the differences in counterfactual outcomes when using the various mitigation strategies compared to the baseline approach, that is, Wachter with $\gamma = 0.5$: choosing a higher decision threshold pushes the counterfactual a little further into the target domain; this effect is even stronger for ClaPROAR; finally, using the Gravitational generator the counterfactual ends up all the way inside the target domain in the neighborhood of $\bar{x}^*$[15]. Linking these ideas back to Example Example 3.2, the mitigation strategies help ensure that the recommended recourse actions are substantial enough to truly lead to an increase in the probability that the admitted student eventually graduates.

Our findings indicate that all three mitigation strategies are at least at par with LS generators with respect to their effectiveness at mitigating domain and model shifts. Figure 3.6 presents a subset of the evaluation metrics for our synthetic data with overlapping classes. The top row in Figure 3.6 indicates that while domain shifts are of roughly the same magnitude for both Wachter and LS generators, our proposed strategies effectively mitigate these shifts. ClaPROAR appears to be particularly

[14] Note that despite the naming conventions, our goal here is not to provide yet more counterfactual generators. Rather than looking at them as isolated entities, we believe and demonstrate that different approaches can be effectively combined.

[15] In order for the Gravitational generator and ClaPROAR to work as expected, one needs to ensure that counterfactual search continues, independent of the threshold probability γ.

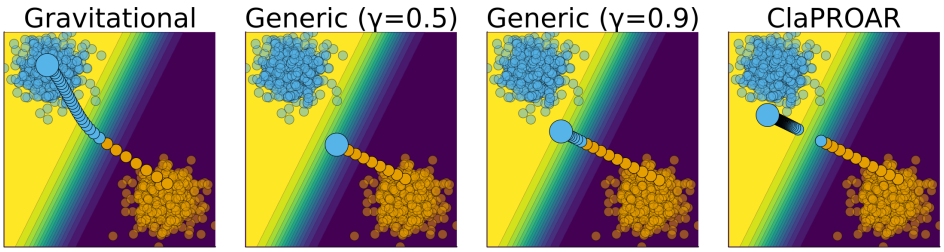


Figure 3.5. Illustrative example demonstrating the properties of the various mitigation strategies. Samples from the negative class ($y = 0$) are marked in orange while samples of the positive class ($y = 1$) are marked in blue.

effective, which is positively surprising since it is designed to explicitly address model shifts, not domain shifts. As evident from the middle row in Figure 3.6 model shifts can also be reduced: for the deep ensemble LS search yields results that are at par with the mitigation strategies, while for both the simple MLP and logistic regression our simple strategies are more effective. The same overall pattern can be observed for out-of-sample model performance. Concerning the other synthetic datasets, for the Moons dataset, the emerging patterns are largely the same, but the estimated model shifts are insignificant as noted earlier; the same holds for the Circles dataset, but there is no significant reduction in model performance for our neural networks; in the case of linearly separable data, we find the Gravitational generator to be most effective at mitigating shifts.

An interesting finding is also that the proposed strategies have a complementary effect when used in combination with LS generators. In experiments we conducted on the synthetic data, the benefits of LS generators were exacerbated further when using a more conservative threshold or combining it with the penalties underlying Gravitational and ClaPROAR. In Figure 3.7 the conventional LS generator with $\gamma = 0.5$ serves as our baseline. Evidently, being more conservative or using one of our proposed penalties decreases the estimated domain and model shifts, in some cases beyond significance.

Finally, Figure 3.8 shows the results for our real-world data. We note that for both the California Housing and GMSC data, ClaPROAR does have an attenuating effect on model performance deterioration[16]. Overall, the results are less significant, possibly because a somewhat smaller share of individuals from the non-target group received recourse than in the synthetic case[17].

[16] Estimated domain shifts (not shown) were largely insubstantial, as in Figure 3.4 in the previous section.

[17] In earlier experiments we moved a larger share of individuals and the results more clearly favoured our mitigation strategies.

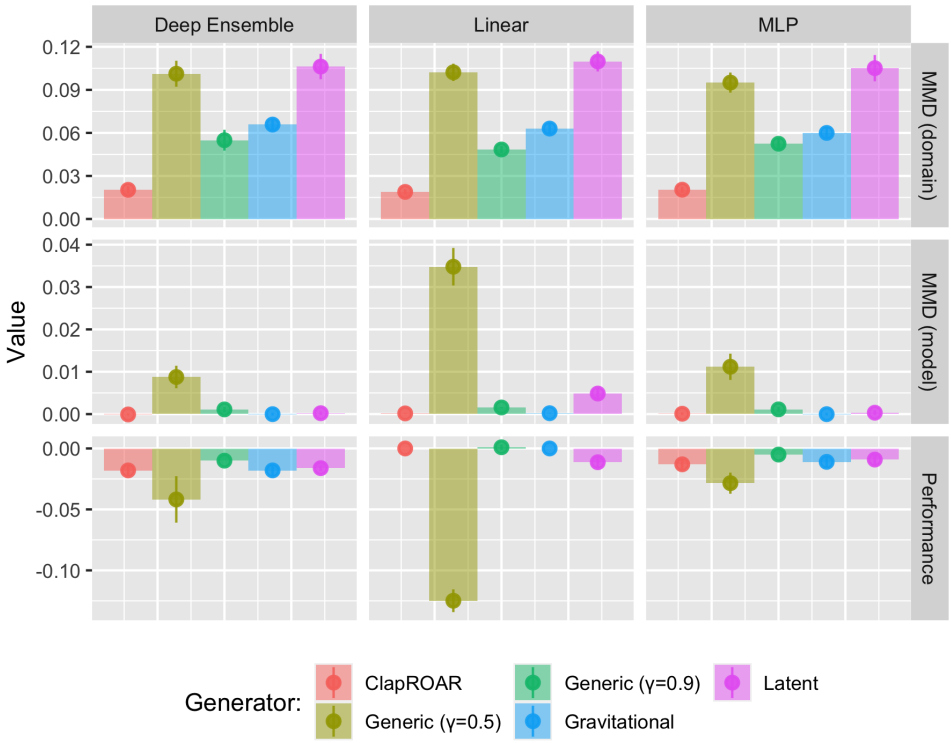


Figure 3.6. The differences in counterfactual outcomes when using the various mitigation strategies compared to the baseline approach, that is Wachter with $\gamma = 0.5$. Results for synthetic data with overlapping classes. The shown model MMD (PP MMD) was computed over a mesh grid of points. Error bars indicate the standard deviation across folds.

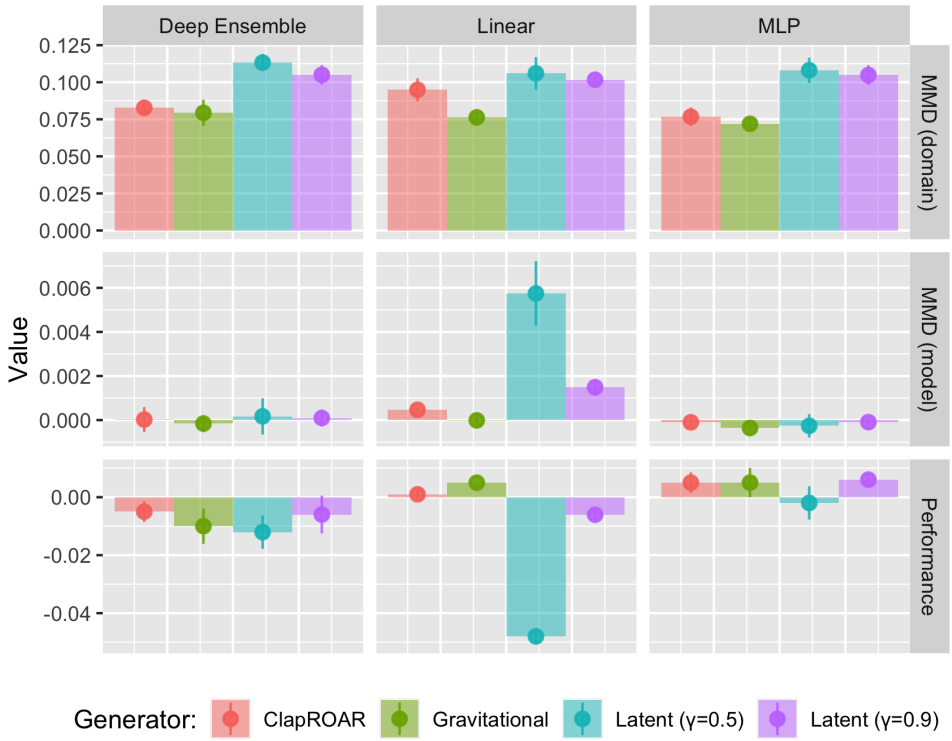


Figure 3.7. Combining various mitigation strategies with LS search. Results for synthetic data with overlapping classes. The shown model MMD (PP MMD) was computed over a mesh grid of points. Error bars indicate the standard deviation across folds.

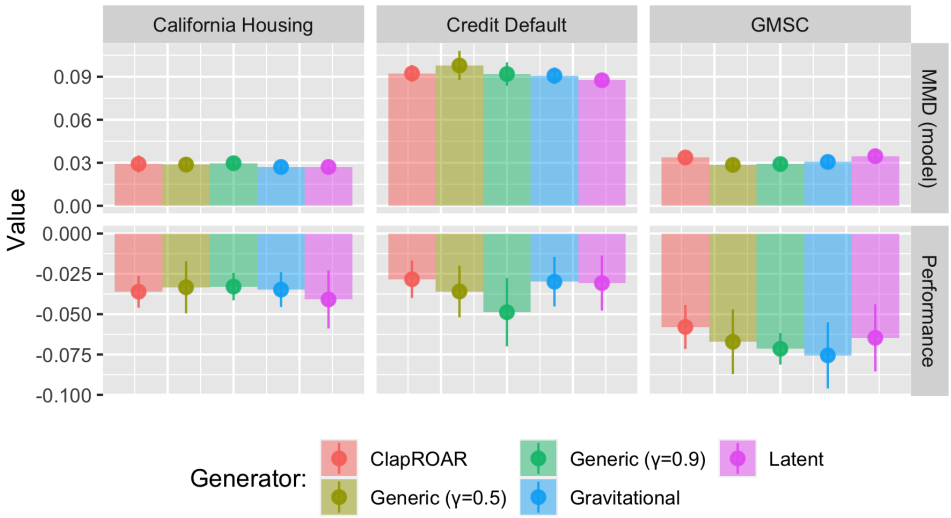


Figure 3.8. The differences in counterfactual outcomes when using the various mitigation strategies compared to the baseline approach, that is Wachter with $\gamma = 0.5$. Results for the MLP using real-world datasets. The shown model MMD (PP MMD) was computed over actual samples, rather than a mesh grid. Error bars indicate the standard deviation across folds.

3.8. DISCUSSION

Our results in Section Section 3.6 indicate that state-of-the-art approaches to Algorithmic Recourse induce substantial domain and model shift if implemented at scale in practice. These induced shifts can and should be considered as an (expected) external cost of individual recourse. While they do not affect the individual directly as long as we look at the individual in isolation, they can be seen to affect the broader group of stakeholders in automated data-driven decision-making. We have seen, for example, that out-of-sample model performance generally deteriorates in our simulation experiments. In practice, this can be seen as a cost to model owners, that is the group of stakeholders using the model as a decision-making tool. As we have set out in Example Example 3.2 of our introduction, these model owners may be unwilling to carry that cost, and hence can be expected to stop offering recourse to individuals altogether. This in turn is costly to those individuals that would otherwise derive utility from being offered recourse.

So, where does this leave us? We would argue that the expected external costs of individual recourse should be shared by all stakeholders. The most straightforward way to achieve this is to introduce a penalty for external costs in the counterfactual search objective function, as we have set out in Equation 3.4. This will on average lead to more costly counterfactual outcomes, but may help to avoid extreme scenarios, in which minimal-cost recourse is reserved to a tiny minority of individuals. We have shown various types of shift-mitigating strategies that can be used to this end. Since all of these strategies can be seen simply as a specific adaption of Equation 3.4, they can be applied to any of the various counterfactual generators studied here.

3.9. LIMITATIONS AND FUTURE WORK

While we believe that this work constitutes a valuable starting point for addressing existing issues in Algorithmic Recourse from a fresh perspective, we are aware of several of its limitations. In the following, we highlight some of these and point to avenues for future research.

3.9.1. PRIVATE VS. EXTERNAL COSTS

Perhaps the most crucial shortcoming of our work is that we merely point out that there exists a trade-off between private costs to the individual and external costs to the collective of stakeholders. We fall short of providing any definitive answers as to how that trade-off may be resolved in practice. The mitigation strategies we have proposed here provide a good starting point, but they are ad-hoc extensions of the existing AR framework. An interesting idea to explore in future work could be the potential for Pareto optimal Algorithmic Recourse, that is, a collective recourse outcome in which no single individual can be made better off, without making at

least one other individual worse off. This type of work would be interdisciplinary and could help to formalize some of the concepts presented in this work.

3.9.2. EXPERIMENTAL SETUP

The experimental setup proposed here is designed to mimic a real-world recourse process in a simple fashion. In practice, models are updated regularly (Upadhyay, Joshi, and Lakkaraju 2021). We also find it plausible to assume that the implementation of recourse happens periodically for different individuals, rather than all at once at time $t = 0$. That being said, our experimental design is a vast over-simplification of potential real-world scenarios. In practice, any endogenous shifts that may occur can be expected to be entangled with exogenous shifts of the nature investigated in Upadhyay, Joshi, and Lakkaraju (2021). We also make implicit assumptions about the utility functions of the involved agents that may well be too simple: individuals seeking recourse are assumed to always implement the proposed Counterfactual Explanations; conversely, the agent in charge of the model M is assumed to always treat individuals that have implemented valid recourse as if they were truly now in the target class.

3.9.3. CAUSAL MODELLING

In this work, we have focused on popular counterfactual generators that do not incorporate any causal knowledge. The generated perturbations therefore may involve changes to variables that affect the outcome predicted by the black-box model, but not the true outcome. The implementation of such changes is typically described as **gaming** (J. Miller, Milli, and Hardt 2020), although they need not be driven by adversarial intentions: in Example Example 3.2, student applicants may dutifully focus on acquiring credentials that help them to be admitted to university, but ultimately not to improve their chances of success at completing their degree (Barocas, Hardt, and Narayanan 2022). Preventing such actions may help to avoid the dynamics we have pointed to in this work. Future work would likely benefit from including recent approaches to AR that incorporate causal knowledge such as Karimi, Schölkopf, and Valera (2021).

3.9.4. CLASSIFIERS

For reasons stated earlier, we have limited our analysis to differentiable linear and non-linear classifiers, in particular logistic regression and deep neural networks. While these sorts of classifiers have also typically been analyzed in the existing literature on Counterfactual Explanations and Algorithmic Recourse, they represent only a subset of popular machine learning models employed in practice. Despite the success and popularity of deep learning in the context of high-dimensional data such as image, audio and video, empirical evidence suggests that other models such as boosted decision trees may have an edge when it comes to lower-dimensional

tabular datasets, such as the ones considered here (Borisov et al. 2022; Grinsztajn, Oyallon, and Varoquaux 2022).

3.9.5. DATA

Largely in line with the existing literature on Algorithmic Recourse, we have limited our analysis of real-world data to three commonly used benchmark datasets that involve binary prediction tasks. Future work may benefit from including novel datasets or extending the analysis to multi-class or regression problems, the latter arguably representing the most common objective in Finance and Economics.

3.10. CONCLUDING REMARKS

This work has revisited and extended some of the most general and defining concepts underlying the literature on Counterfactual Explanations and, in particular, Algorithmic Recourse. We demonstrate that long-held beliefs as to what defines optimality in AR, may not always be suitable. Specifically, we run experiments that simulate the application of recourse in practice using various state-of-the-art counterfactual generators and find that all of them induce substantial domain and model shifts. We argue that these shifts should be considered as an expected external cost of individual recourse and call for a paradigm shift from individual to collective recourse in these types of situations. By proposing an adapted counterfactual search objective that incorporates this cost, we make that paradigm shift explicit. We show that this modified objective lends itself to mitigation strategies that can be used to effectively decrease the magnitude of induced domain and model shifts. Through our work, we hope to inspire future research on this important topic. To this end we have open-sourced all of our code along with a Julia package: `AlgorithmicRecourseDynamics.jl`. Future researchers should find it easy to replicate, modify and extend the simulation experiments presented here and apply them to their own custom counterfactual generators.

3.11. ACKNOWLEDGEMENTS

Some of the members of TU Delft were partially funded by ICAI AI for Fintech Research, an ING — TU Delft collaboration.

4

FAITHFUL MODEL EXPLANATIONS THROUGH ENERGY-CONSTRAINED CONFORMAL COUNTERFACTUALS

Counterfactual explanations offer an intuitive and straightforward way to explain black-box models and offer algorithmic recourse to individuals. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic explanations for the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily describe the behavior of the black-box model faithfully. We formalize this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **E**nergy-**C**onstrained **C**onformal **Co**unterfactuals that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that *ECCCo* reconciles the need for faithfulness and plausibility. In particular, we show that for models with gradient access, it is possible to achieve state-of-the-art performance without the need for surrogate models. To do so, our framework relies solely on properties defining the black-box model itself by leveraging recent advances in energy-based modelling and conformal prediction. To our knowledge, this is the first venture in this direction for generating faithful counterfactual explanations. Thus, we anticipate that *ECCCo* can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

This chapter was published in Proceedings of the AAAI Conference on Artificial Intelligence by Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem (2024a). See Chapter 1.8 for additional publication details.

4.1. INTRODUCTION

Counterfactual explanations provide a powerful, flexible and intuitive way to not only explain black-box models but also offer the possibility of algorithmic recourse to affected individuals. Instead of opening the black box, counterfactual explanations work under the premise of strategically perturbing model inputs to understand model behavior (Wachter, Mittelstadt, and Russell 2017). Intuitively speaking, we generate explanations in this context by asking what-if questions of the following nature: 'Our credit risk model currently predicts that this individual is not creditworthy. What if they reduced their monthly expenditures by 10%?'

This is typically implemented by defining a target outcome $\mathbf{y}^+ \in \mathcal{Y}$ for some individual $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ described by D attributes, for which the model $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ initially predicts a different outcome: $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$. Counterfactuals are then searched by minimizing a loss function that compares the predicted model output to the target outcome: $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^+)$. Since counterfactual explanations work directly with the black-box model, valid counterfactuals always have full local fidelity by construction where fidelity is defined as the degree to which explanations approximate the predictions of a black-box model (Molnar 2022).

In situations where full fidelity is a requirement, counterfactual explanations offer a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), which involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring that an explanation *faithfully* describes the behavior of a model. That is because multiple distinct explanations can lead to the same model prediction, especially when dealing with heavily parameterized models like deep neural networks, which are underspecified by the data (Wilson 2020). In the context of counterfactuals, the idea that no two explanations are the same arises almost naturally. A key focus in the literature has therefore been to identify those explanations that are most appropriate based on a myriad of desiderata such as closeness (Wachter, Mittelstadt, and Russell 2017), sparsity (Schut et al. 2021), actionability (Ustun, Spangher, and Liu 2019) and plausibility (Joshi et al. 2019).

In this work, we draw closer attention to modelling faithfulness rather than fidelity as a desideratum for counterfactuals. We define faithfulness as the degree to which counterfactuals are consistent with what the model has learned about the data. Our

key contributions are as follows: first, we show that fidelity is an insufficient evaluation metric for counterfactuals (Section 4.3) and propose a definition of faithfulness that gives rise to more suitable metrics (Section 4.4). Next, we introduce a *ECCCo*: a novel algorithmic approach aimed at generating energy-constrained conformal counterfactuals that faithfully explain model behavior in Section 4.5. Finally, we provide extensive empirical evidence demonstrating that *ECCCo* faithfully explains model behavior and attains plausibility only when appropriate (Section 4.6).

To our knowledge, this is the first venture in this direction for generating faithful counterfactuals. Thus, we anticipate that *ECCCo* can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

4.2. BACKGROUND

While counterfactual explanations (CE) can also be generated for arbitrary regression models (Spooner et al. 2021), existing work has primarily focused on classification problems. Let $\mathcal{Y} = (0,1)^K$ denote the one-hot-encoded output domain with K classes. Then most counterfactual generators rely on gradient descent to optimize different flavors of the following counterfactual search objective:

$$\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda \text{cost}(f(\mathbf{Z}'))\} \tag{4.1}$$

Here $\text{yloss}(\cdot)$ denotes the primary loss function, $f(\cdot)$ is a function that maps from the counterfactual state space to the feature space and $\text{cost}(\cdot)$ is either a single penalty or a collection of penalties that are used to impose constraints through regularization. Equation 5.1 restates the baseline approach to gradient-based counterfactual search proposed by Wachter, Mittelstadt, and Russell (2017) in general form as introduced Chapter 2. To explicitly account for the multiplicity of explanations, $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes an L-dimensional array of counterfactual states.

The baseline approach, which we will simply refer to as *Wachter*, searches a single counterfactual directly in the feature space and penalizes its distance to the original factual. In this case, $f(\cdot)$ is simply the identity function and $\mathcal{Z}$ corresponds to the feature space itself. Many derivative works of Wachter, Mittelstadt, and Russell (2017) have proposed new flavors of Equation 5.1, each of them designed to address specific *desiderata* that counterfactuals ought to meet in order to properly serve both AI practitioners and individuals affected by algorithmic decision-making systems. The list of desiderata includes but is not limited to the following: sparsity, closeness (Wachter, Mittelstadt, and Russell 2017), actionability (Ustun, Spangher, and Liu 2019), diversity (Mothilal, Sharma, and Tan 2020), plausibility (Joshi et al. 2019; Poyiadzi et al. 2020; Schut et al. 2021), robustness (Upadhyay, Joshi, and Lakkaraju 2021; Pawelczyk et al. 2023; Altmeyer, Angela, et al. 2023) and causality (Karimi, Schölkopf, and Valera 2021). Different counterfactual generators addressing these needs have been extensively surveyed and evaluated in various

studies (Verma et al. 2022; Karimi et al. 2021; Pawelczyk et al. 2021; Artelt et al. 2021; Guidotti 2022).

The notion of plausibility is central to all of the desiderata. For example, Artelt et al. (2021) find that plausibility typically also leads to improved robustness. Similarly, plausibility has also been connected to causality in the sense that plausible counterfactuals respect causal relationships (Mahajan, Tan, and Sharma 2020). Consequently, the plausibility of counterfactuals has been among the primary concerns for researchers. Achieving plausibility is equivalent to ensuring that the generated counterfactuals comply with the true and unobserved data-generating process (DGP). We define plausibility formally in this work as follows:

Definition 4.1 (Plausible Counterfactuals)**.** Let $\mathcal{X}|\mathbf{y}^+ = p(\mathbf{x}|\mathbf{y}^+)$ denote the true conditional distribution of samples in the target class $\mathbf{y}^+$. Then for $\mathbf{x}'$ to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.

To generate plausible counterfactuals, we first need to quantify the conditional distribution of samples in the target class ($\mathcal{X}|\mathbf{y}^+$). We can then ensure that we generate counterfactuals that comply with that distribution.

One straightforward way to do this is to use surrogate models for the task. Joshi et al. (2019), for example, suggest that instead of searching counterfactuals in the feature space $\mathcal{X}$, we can traverse a latent embedding $\mathcal{Z}$ (Equation 5.1) that implicitly codifies the DGP. To learn the latent embedding, they propose using a generative model such as a Variational Autoencoder (VAE). Provided the surrogate model is well-specified, their proposed approach *REVISE* can yield plausible explanations. Others have proposed similar approaches: Dombrowski, Gerken, and Kessel (2021) traverse the base space of a normalizing flow to solve Equation 5.1; Poyiadzi et al. (2020) use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to constrain the counterfactuals to dense regions in the feature space; finally, Karimi, Schölkopf, and Valera (2021) assume knowledge about the causal graph that generates the data.

A competing approach towards plausibility that is also closely related to this work instead relies on the black-box model itself. Schut et al. (2021) show that to meet the plausibility objective we need not explicitly model the input distribution. Pointing to the undesirable engineering overhead induced by surrogate models, they propose to rely on the implicit minimization of predictive uncertainty instead. Their proposed methodology, which we will refer to as *Schut*, solves Equation 5.1 by greedily applying Jacobian-Based Saliency Map Attacks (JSMA) in the feature space with cross-entropy loss and no penalty at all. The authors demonstrate theoretically and empirically that their approach yields counterfactuals for which the model M_θ predicts the target label $\mathbf{y}^+$ with high confidence. Provided the model is well-specified, these counterfactuals are plausible. This idea hinges on the assumption that the black-box model provides well-calibrated predictive uncertainty estimates.

4.3. WHY FIDELITY IS NOT ENOUGH: A MOTIVATIONAL EXAMPLE

As discussed in the introduction, any valid counterfactual also has full fidelity by construction: solutions to Equation 5.1 are considered valid as soon as the label predicted by the model matches the target class. So while fidelity always applies, counterfactuals that address the various desiderata introduced above can look vastly different from each other.

To demonstrate this with an example, we have trained a simple image classifier M_θ on the well-known *MNIST* dataset (LeCun et al. 1998): a Multi-Layer Perceptron (*MLP*) with test set accuracy > 0.9. No measures have been taken to improve the model's adversarial robustness or its capacity for predictive uncertainty quantification. The far left panel of Figure 4.1 shows a random sample drawn from the dataset. The underlying classifier correctly predicts the label 'nine' for this image. For the given factual image and model, we have used *Wachter*, *Schut* and *REVISE* to generate one counterfactual each in the target class 'seven'. The perturbed images are shown next to the factual image from left to right in Figure 4.1. Captions on top of the images indicate the generator along with the predicted probability that the image belongs to the target class. In all cases, that probability is very high, while the counterfactuals look very different.

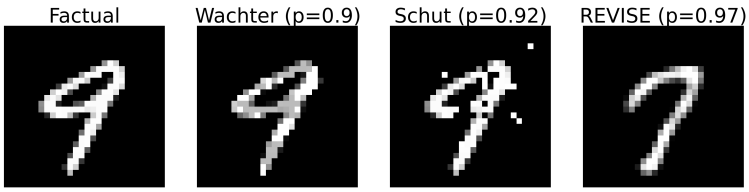


Figure 4.1. Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left), then the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

Since *Wachter* is only concerned with closeness, the generated counterfactual is almost indistinguishable from the factual. *Schut* expects a well-calibrated model that can generate predictive uncertainty estimates. Since this is not the case, the generated counterfactual looks like an adversarial example. Finally, the counterfactual generated by *REVISE* looks much more plausible than the other two. But is it also more faithful to the behavior of our *MNIST* classifier? That is much less clear because the surrogate used by *REVISE* introduces friction: explanations no longer depend exclusively on the black-box model itself.

So which of the counterfactuals most faithfully explains the behavior of our image classifier? Fidelity cannot help us to make that judgement, because all of these counterfactuals have full fidelity. Thus, fidelity is an insufficient evaluation metric to assess the faithfulness of CE.

4.4. FAITHFUL FIRST, PLAUSIBLE SECOND

Considering the limitations of fidelity as demonstrated in the previous section, analogous to Definition 4.1, we introduce a new notion of faithfulness in the context of CE:

Definition 4.2 (Faithful Counterfactuals)**.** Let $\mathcal{X}_\theta|\mathbf{y}^+ = p_\theta(\mathbf{x}|\mathbf{y}^+)$ denote the conditional distribution of $\mathbf{x}$ in the target class $\mathbf{y}^+$, where θ denotes the parameters of model M_θ. Then for $\mathbf{x}'$ to be considered a faithful counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^+$.

In doing this, we merge in and nuance the concept of plausibility (Definition 4.1) where the notion of 'consistent with the data' becomes 'consistent with what the model has learned about the data'.

4.4.1. QUANTIFYING THE MODEL'S GENERATIVE PROPERTY

To assess counterfactuals with respect to Definition 4.2, we need a way to quantify the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}^+)$. To this end, we draw on ideas from energy-based modelling (EBM), a subdomain of machine learning that is concerned with generative or hybrid modelling (Grathwohl et al. 2020; Du and Mordatch 2020). In particular, note that if we fix $\mathbf{y}$ to our target value $\mathbf{y}^+$, we can conditionally draw from $p_\theta(\mathbf{x}|\mathbf{y}^+)$ by randomly initializing $\mathbf{x}_0$ and then using Stochastic Gradient Langevin Dynamics (SGLD) as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon_j^2}{2}\mathcal{E}_\theta(\mathbf{x}_j|\mathbf{y}^+) + \epsilon_j \mathbf{r}_j, \quad j = 1, ..., J \tag{4.2}$$

where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ_j is typically polynomially decayed (Welling and Teh 2011). The term $\mathcal{E}_\theta(\mathbf{x}_j|\mathbf{y}^+)$ denotes the model energy conditioned on the target class label $\mathbf{y}^+$ which we specify as the negative logit corresponding to $\mathbf{y}^+$. To allow for faster sampling, we follow the common practice of choosing the step-size ϵ_j and the standard deviation of $\mathbf{r}_j$ separately. While $\mathbf{x}_J$ is only guaranteed to distribute as $p_\theta(\mathbf{x}|\mathbf{y}^+)$ if $\epsilon \to 0$ and $J \to \infty$, the bias introduced for a small finite ϵ is negligible in practice (Murphy 2023).

Generating multiple samples using SGLD thus yields an empirical distribution $\widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}$ that approximates what the model has learned about the input data. While in the context of EBM, this is usually done during training, we propose to repurpose this approach during inference in order to evaluate the faithfulness of model explanations. The appendix provides additional implementation details for any tasks related to energy-based modelling[1].

[1]The supplementary appendix can be found here: https://arxiv.org/abs/2312.10648.

4.4.2. Quantifying the model's predictive uncertainty

Faithful counterfactuals can be expected to also be plausible if the learned conditional distribution $\mathcal{X}_\theta|\mathbf{y}^+$ (Definition 4.2) is close to the true conditional distribution $\mathcal{X}|\mathbf{y}^+$ (Definition 4.1). We can further improve the plausibility of counterfactuals without the need for surrogate models that may interfere with faithfulness by minimizing predictive uncertainty (Schut et al. 2021). Unfortunately, this idea relies on the assumption that the model itself provides predictive uncertainty estimates, which may be too restrictive in practice.

To relax this assumption, we use conformal prediction (CP), an approach to predictive uncertainty quantification that has recently gained popularity (Angelopoulos and Bates 2022; Manokhin 2022). Crucially for our intended application, CP is model-agnostic and can be applied during inference without placing any restrictions on model training. It works under the premise of turning heuristic notions of uncertainty into rigorous estimates by repeatedly sifting through the training data or a dedicated calibration dataset. Calibration data is used to compute so-called nonconformity scores: $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ where $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function* (see appendix for details).

Conformal classifiers produce prediction sets for individual inputs that include all output labels that can be reasonably attributed to the input. These sets are formed as follows,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \tag{4.3}$$

where $\hat{q}$ denotes the $(1-\alpha)$-quantile of $\mathcal{S}$ and α is a predetermined error rate. These sets tend to be larger for inputs that do not conform with the training data and are characterized by high predictive uncertainty. To leverage this notion of predictive uncertainty in the context of gradient-based counterfactual search, we use a smooth set size penalty introduced by Stutz et al. (2022):

$$\Omega(C_\theta(\mathbf{x}; \alpha)) = \max\left(0, \sum_{\mathbf{y} \in \mathcal{Y}} C_{\theta,\mathbf{y}}(\mathbf{x}_i; \alpha) - \kappa\right) \tag{4.4}$$

Here, $\kappa \in \{0, 1\}$ is a hyper-parameter and $C_{\theta,\mathbf{y}}(\mathbf{x}_i; \alpha)$ can be interpreted as the probability of label $\mathbf{y}$ being included in the prediction set (see appendix for details). In order to compute this penalty for any black-box model, we merely need to perform a single calibration pass through a holdout set $\mathcal{D}_{\text{cal}}$. Arguably, data is typically abundant and in most applications, practitioners tend to hold out a test data set anyway. Consequently, CP removes the restriction on the family of predictive models, at the small cost of reserving a subset of the available data for calibration. This particular case of conformal prediction is referred to as *split conformal prediction* (SCP) as it involves splitting the training data into a proper training dataset and a calibration dataset.

4.4.3. EVALUATING PLAUSIBILITY AND FAITHFULNESS

The parallels between our definitions of plausibility and faithfulness imply that we can also use similar evaluation metrics in both cases. Since existing work has focused heavily on plausibility, it offers a useful starting point. In particular, Guidotti (2022) have proposed an implausibility metric that measures the distance of the counterfactual from its nearest neighbor in the target class. As this distance is reduced, counterfactuals get more plausible under the assumption that the nearest neighbor itself is plausible in the sense of Definition 4.1. In this work, we use the following adapted implausibility metric,

$$\text{impl}(\mathbf{x}', \mathbf{X}_{\mathbf{y}^+}) = \frac{1}{|\mathbf{X}_{\mathbf{y}^+}|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{4.5}$$

where $\mathbf{x}'$ denotes the counterfactual and $\mathbf{X}_{\mathbf{y}^+}$ is a subsample of the training data in the target class $\mathbf{y}^+$. By averaging over multiple samples in this manner, we avoid the risk that the nearest neighbor of $\mathbf{x}'$ itself is not plausible according to Definition 4.1 (e.g. an outlier).

Equation 4.5 gives rise to a similar evaluation metric for unfaithfulness. We swap out the subsample of observed individuals in the target class for the set of samples generated through SGLD ($\widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}$):

$$\text{unfaith}(\mathbf{x}', \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}) = \frac{1}{|\widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}|} \sum_{\mathbf{x} \in \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{4.6}$$

Our default choice for the dist$(\cdot)$ function in both cases is the Euclidean Norm. Depending on the type of input data other choices may be more adequate (see Section 4.6.1).

4.5. ENERGY-CONSTRAINED CONFORMAL COUNTERFACTUALS

Given our proposed notion of faithfulness, we now describe *ECCCo*, our proposed framework for generating Energy-Constrained Conformal Counterfactuals. It is based on the premise that counterfactuals should first and foremost be faithful. Plausibility, as a secondary concern, is then still attainable to the degree that the black-box model itself has learned plausible explanations for the underlying data.

We begin by substituting the loss function in Equation 5.1,

$$\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{L_{\text{JEM}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+) + \lambda \text{cost}(f(\mathbf{Z}'))\} \tag{4.7}$$

where $L_{\text{JEM}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+)$ is a hybrid loss function used in joint-energy modelling evaluated at a given counterfactual state for a given model and target outcome:

$$L_{\text{JEM}}(f(\mathbf{Z}'); \cdot) = L_{\text{clf}}(f(\mathbf{Z}'); \cdot) + L_{\text{gen}}(f(\mathbf{Z}'); \cdot) \tag{4.8}$$

The first term, L_{clf}, is any standard classification loss function such as cross-entropy loss. The second term, L_{gen}, is used to measure loss with respect to the generative task[2]. In the context of joint-energy training, L_{gen} induces changes in model parameters θ that decrease the energy of observed samples and increase the energy of samples generated through SGLD (Du and Mordatch 2020).

The key observation in our context is that we can rely solely on decreasing the energy of the counterfactual itself. This is sufficient to capture the generative property of the underlying model since it is implicitly captured by its parameters θ. Importantly, this means that we do not need to generate conditional samples through SGLD during our counterfactual search at all (see appendix for details).

This observation leads to the following simple objective function for *ECCCo*:

$$\begin{aligned} &\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{L_{\text{clf}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+) + \lambda_1 \text{cost}(f(\mathbf{Z}')) \\ &+ \lambda_2 \mathcal{E}_\theta(f(\mathbf{Z}')|\mathbf{y}^+) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha))\} \end{aligned} \tag{4.9}$$

The first penalty term involving λ_1 induces closeness like in Wachter, Mittelstadt, and Russell (2017). The second penalty term involving λ_2 induces faithfulness by constraining the energy of the generated counterfactual. The third and final penalty term involving λ_3 ensures that the generated counterfactual is associated with low predictive uncertainty. To tune these hyperparameters we have relied on grid search.

Concerning feature autoencoding ($f : \mathcal{Z} \mapsto \mathcal{X}$), *ECCCo* does not rely on latent space search to achieve its primary objective of faithfulness. By default, we choose $f(\cdot)$ to be the identity function as in *Wachter*. This is generally also enough to achieve plausibility, provided the model has learned plausible explanations for the data. In some cases, plausibility can be improved further by mapping counterfactuals to a lower-dimensional latent space. In the following, we refer to this approach as *ECCCo+*: that is, *ECCCo* plus dimensionality reduction.

Figure 5.1 illustrates how the different components in Equation 4.9 affect the counterfactual search for a synthetic dataset. The underlying classifier is a Joint Energy Model (*JEM*) that was trained to predict the output class (blue or orange) and generate class-conditional samples (Grathwohl et al. 2020). We have used four different generator flavors to produce a counterfactual in the blue class for a sample from the orange class: *Wachter*, which only uses the first penalty ($\lambda_2 = \lambda_3 = 0$);

[2] In practice, regularization loss is typically also added. We follow this convention but have omitted the term here for simplicity.

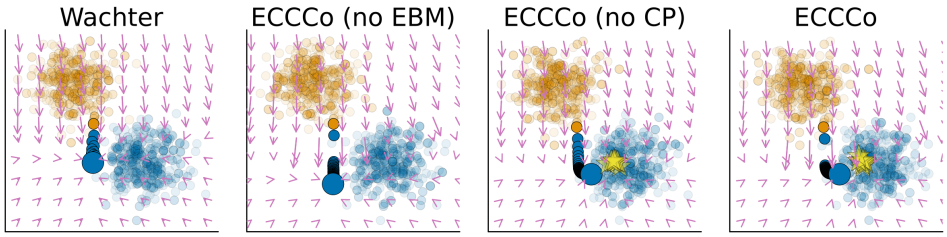


Figure 4.2. Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfactual in the blue class for a sample from the orange class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

ECCCo (no EBM), which does not constrain energy ($\lambda_2 = 0$); *ECCCo (no CP)*, which involves no set size penalty ($\lambda_3 = 0$); and, finally, *ECCCo*, which involves all penalties defined in Equation 4.9. Arrows indicate (negative) gradients with respect to the objective function at different points in the feature space.

While *Wachter* generates a valid counterfactual, it ends up close to the original starting point consistent with its objective. *ECCCo (no EBM)* avoids regions of high predictive uncertainty near the decision boundary, but the outcome is still not plausible. The counterfactual produced by *ECCCo (no CP)* is energy-constrained. Since the *JEM* has learned the conditional input distribution reasonably well in this case, the counterfactual is both faithful and plausible. Finally, the outcome for *ECCCo* looks similar, but the additional smooth set size penalty leads to somewhat faster convergence.

4.6. EMPIRICAL ANALYSIS

Our goal in this section is to shed light on the following research questions:

Research Question 4.1 (Faithfulness)**.** To what extent are counterfactuals generated by *ECCCo* more faithful than those produced by state-of-the-art generators?

Research Question 4.2 (Balancing Desiderata)**.** Compared to state-of-the-art generators, how does *ECCCo* balance the two key objectives of faithfulness and plausibility?

The second question is motivated by the intuition that faithfulness and plausibility should coincide for models that have learned plausible explanations of the data.

4.6.1. Experimental Setup

To assess and benchmark the performance of our proposed generator against the state of the art, we generate multiple counterfactuals for different models and datasets. In particular, we compare *ECCCo* and its variants to the following counterfactual generators that were introduced above: firstly, *Schut*, which works under the premise of minimizing predictive uncertainty; secondly, *REVISE*, which is state-of-the-art (SOTA) with respect to plausibility; and, finally, *Wachter*, which serves as our baseline. In the case of *ECCCo+*, we use principal component analysis (PCA) for dimensionality reduction: the latent space $\mathcal{Z}$ is spanned by the first n_z principal components where we choose n_z to be equal to the latent dimension of the VAE used by *REVISE*.

For the predictive modelling tasks, we use multi-layer perceptrons (*MLP*), deep ensembles, joint energy models (*JEM*) and convolutional neural networks (LeNet-5 *CNN* (LeCun et al. 1998)). Both joint-energy modelling and ensembling have been associated with improved generative properties and adversarial robustness (Grathwohl et al. 2020; Lakshminarayanan, Pritzel, and Blundell 2017), so we expect this to be positively correlated with the plausibility of *ECCCo*. To account for stochasticity, we generate many counterfactuals for each target class, generator, model and dataset over multiple runs.

We perform benchmarks on eight datasets from different domains. From the credit and finance domain we include three tabular datasets: Give Me Some Credit (*GMSC*) (Kaggle 2011), *German Credit* (Hoffman 1994) and *California Housing* (Pace and Barry 1997). All of these are commonly used in the related literature (Karimi et al. 2021; Altmeyer, Angela, et al. 2023; Pawelczyk et al. 2021). Following related literature (Schut et al. 2021; Dhurandhar et al. 2018) we also include two image datasets: *MNIST* (LeCun et al. 1998) and *Fashion MNIST* (Xiao, Rasul, and Vollgraf 2017).

Full details concerning model training as well as detailed descriptions and results for all datasets can be found in the appendix. In the following, we will focus on the most relevant results highlighted in Table 4.1 and Table 4.2. The tables show sample averages along with standard deviations across multiple runs for our key evaluation metrics for the *California Housing* and *GMSC* datasets (Table 4.1) and the *MNIST* dataset (Table 4.2). For each metric, the best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (Wachter). For the tabular datasets, we use the default Euclidean distance to measure unfaithfulness and implausibility as defined in Equation 4.6 and Equation 4.5, respectively. The third metric presented in Table 4.1 quantifies the predictive uncertainty of the counterfactual as measured by Equation 4.4. For the vision datasets, we rely on measuring the structural dissimilarity between images for our unfaithfulness and implausibility metrics (Wang, Simoncelli, and Bovik 2003).

4.6.2. FAITHFULNESS

Table 4.1. Results for tabular datasets: sample averages +/- one standard deviation across valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*)

		California Housing			GMSC		
Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓
MLP Ensemble	ECCCo	**3.69 ± 0.08****	1.94 ± 0.13	**0.09 ± 0.01****	3.84 ± 0.07**	2.13 ± 0.08	**0.23 ± 0.01****
	ECCCo+	3.88 ± 0.07**	1.20 ± 0.09	0.15 ± 0.02	**3.79 ± 0.05****	1.81 ± 0.05	0.30 ± 0.01*
	ECCCo (no CP)	3.70 ± 0.08**	1.94 ± 0.13	0.10 ± 0.01**	3.85 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**
	ECCCo (no EBM)	4.03 ± 0.07	1.12 ± 0.12	0.14 ± 0.01**	4.08 ± 0.06	0.97 ± 0.08	0.31 ± 0.01*
	REVISE	3.96 ± 0.07*	**0.58 ± 0.03****	0.17 ± 0.03	4.09 ± 0.07	**0.63 ± 0.02****	0.33 ± 0.06
	Schut	4.00 ± 0.06	1.15 ± 0.12	0.10 ± 0.01**	4.04 ± 0.08	1.21 ± 0.08	0.30 ± 0.01*
	Wachter	4.04 ± 0.07	1.13 ± 0.12	0.16 ± 0.01	4.10 ± 0.07	0.95 ± 0.08	0.32 ± 0.01
JEM Ensemble	ECCCo	1.40 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	1.20 ± 0.06*	0.78 ± 0.07**	0.38 ± 0.01
	ECCCo+	**1.28 ± 0.08****	0.60 ± 0.04**	0.11 ± 0.00**	**1.01 ± 0.07****	0.70 ± 0.07**	0.37 ± 0.01
	ECCCo (no CP)	1.39 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	1.21 ± 0.07*	0.77 ± 0.07**	0.39 ± 0.01
	ECCCo (no EBM)	1.70 ± 0.09	0.99 ± 0.08	0.14 ± 0.00*	1.31 ± 0.07	0.97 ± 0.10	0.32 ± 0.01**
	REVISE	1.39 ± 0.15**	**0.59 ± 0.04****	0.25 ± 0.07	1.01 ± 0.07**	**0.63 ± 0.04****	0.33 ± 0.07
	Schut	1.59 ± 0.10*	1.10 ± 0.06	**0.09 ± 0.00****	1.34 ± 0.07	1.21 ± 0.10	**0.26 ± 0.01****
	Wachter	1.71 ± 0.09	0.99 ± 0.08	0.14 ± 0.00	1.31 ± 0.08	0.95 ± 0.10	0.33 ± 0.01

Overall, we find strong empirical evidence suggesting that *ECCCo* consistently achieves state-of-the-art faithfulness. Across all models and datasets highlighted here, different variations of *ECCCo* consistently outperform other generators with respect to faithfulness, in many cases substantially. This pattern is mostly robust across all other datasets.

In particular, we note that the best results are generally obtained when using the full *ECCCo* objective (Equation 4.9). In other words, constraining both energy and predictive uncertainty typically yields the most faithful counterfactuals. We expected the former to play a more significant role in this context and that is typically what we find across all datasets. The results in Table 4.1 indicate that faithfulness can be improved substantially by relying solely on the energy constraint (*ECCCo (no CP)*). In most cases, however, the full objective yields the most faithful counterfactuals. This indicates that predictive uncertainty minimization plays an important role in achieving faithfulness.

We also generally find that latent space search does not impede faithfulness for *ECCCo*. In most cases *ECCCo+* is either on par with *ECCCo* or even outperforms it. There are some notable exceptions though. Cases in which *ECCCo* achieves substantially better faithfulness without latent space search tend to involve more vulnerable models like the simple MLP for MNIST in Table 4.2. We explain this finding as follows: even though dimensionality reduction through PCA in the case of *ECCCo+* can be considered a relatively mild form of intervention, the first n_z principal components fail to capture some of the variation in the data. More vulnerable models may be particularly sensitive to this residual variation in the data.

Consistent with this finding, we also observe that *REVISE* ranks higher for faithfulness, if the model itself has learned more plausible representations of the underlying

data: *REVISE* generates more faithful counterfactuals than the baseline for the *JEM* Ensemble in Table 4.1 and the LeNet-5 *CNN* in Table 4.2. This demonstrates that the two desiderata—faithfulness and plausibility—are not mutually exclusive.

4.6.3. BALANCING DESIDERATA

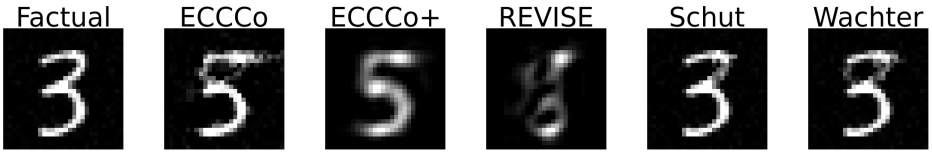


Figure 4.3. Counterfactuals for turning a 3 into a 5: factual (left), then the counterfactuals generated by *ECCCo*, *ECCCo+*, *REVISE*, *Schut* and *Wachter*.

Table 4.2. Results for vision dataset. Formatting details are the same as in Table 4.1.

		MNIST	
Model	Generator	Unfaithfulness ↓	Implausibility ↓
MLP	ECCCo	**0.243 ± 0.000****	0.420 ± 0.001
	ECCCo+	0.246 ± 0.000*	0.306 ± 0.001**
	REVISE	0.248 ± 0.000	**0.301 ± 0.004****
	Schut	0.247 ± 0.001	0.303 ± 0.008**
	Wachter	0.247 ± 0.000	0.344 ± 0.002
LeNet-5	ECCCo	0.248 ± 0.000**	0.387 ± 0.002
	ECCCo+	**0.248 ± 0.000****	0.310 ± 0.002**
	REVISE	0.248 ± 0.000**	0.301 ± 0.002**
	Schut	0.250 ± 0.002	**0.289 ± 0.024***
	Wachter	0.249 ± 0.000	0.335 ± 0.002

Overall, we find strong empirical evidence suggesting that *ECCCo* can achieve near state-of-the-art plausibility without sacrificing faithfulness. Figure 4.3 shows one such example taken from the *MNIST* benchmark where the objective is to turn the factual 'three' (far left) into a 'five'. The underlying model is a LeNet-5 *CNN*. The different images show the counterfactuals produced by the generators, of which

all but the one produced by *Schut* are valid. Both variations of *ECCCo* produce plausible counterfactuals.

Looking at the benchmark results presented in Table 4.1 and Table 4.2 we firstly note that although *REVISE* generally performs best, *ECCCo* and in particular *ECCCo+* often approach SOTA performance. Upon visual inspection of the generated images we actually find that *ECCCo+* performs much better than *REVISE* (see appendix). Zooming in on the details we observe that *ECCCo* and its variations do particularly well, whenever the underlying model has been explicitly trained to learn plausible representations of the data. For both tabular datasets in Table 4.1, *ECCCo* improves plausibility substantially compared to the baseline. This broad pattern is mostly consistent for all other datasets, although there are notable exceptions for which *ECCCo* takes the lead on both plausibility and faithfulness.

While we maintain that generally speaking plausibility should hinge on the quality of the model, our results also indicate that it is possible to balance faithfulness and plausibility if needed: *ECCCo+* generally outperforms other variants of *ECCCo* in this context, occasionally at the small cost of slightly reduced faithfulness. For the vision datasets especially, we find that *ECCCo+* is consistently second only to *REVISE* for all models and regularly substantially better than the baseline. Looking at the *California Housing* data, latent space search markedly improves plausibility without sacrificing faithfulness: for the *JEM* Ensemble, *ECCCo+* performs substantially better than the baseline and only marginally worse than *REVISE*. Importantly, *ECCCo+* does not attain plausibility at all costs: for the *MLP* Ensemble, plausibility is still very low, but this seems to faithfully represent what the model has learned.

We conclude from the findings presented thus far that *ECCCo* enables us to reconcile the objectives of faithfulness and plausibility. It produces plausible counterfactuals if and only if the model itself has learned plausible explanations for the data. It thus avoids the risk of generating plausible but potentially misleading explanations for models that are highly susceptible to implausible explanations.

4.6.4. Additional Desiderata

While we have deliberately focused on our key metrics of interest so far, it is worth briefly considering other common desiderata for counterfactuals. With reference to the right-most columns for each dataset in Table 4.1, we firstly note that *ECCCo* typically reduces predictive uncertainty as intended. Consistent with its design, *Schut* performs well on this metric even though it does not explicitly address uncertainty as measured by conformal prediction set sizes.

Another commonly discussed desideratum is closeness (Wachter, Mittelstadt, and Russell 2017): counterfactuals that are closer to their factuals are associated with smaller costs to individuals in the context of algorithmic recourse. As evident from the additional tables in the appendix, the closeness desideratum tends to be negatively correlated with plausibility and faithfulness. Consequently, both *REVISE* and

ECCCo generally yield more costly counterfactuals than the baseline. Nonetheless, *ECCCo* does not seem to stretch costs unnecessarily: in Figure 4.3 useful parts of the factual 'three' are clearly retained.

4.7. LIMITATIONS

Despite having taken considerable measures to study our methodology carefully, limitations can still be identified.

Firstly, we recognize that our proposed distance-based evaluation metrics for plausibility and faithfulness may not be universally applicable to all types of data. In any case, they depend on choosing a distance metric on a case-by-case basis, as we have done in this work. Arguably, commonly used metrics for measuring other desiderata such as closeness suffer from the same pitfall. We therefore think that future work on counterfactual explanations could benefit from defining universal evaluation metrics.

Relatedly, we note that our proposed metric for measuring faithfulness depends on the availability of samples generated through SGLD, which in turn requires gradient access for models. This means it cannot be used to evaluate non-differentiable classifiers. Consequently, we also have not applied *ECCCo* to some machine learning models commonly used for classification such as decision trees. Since *ECCCo* itself does not rely on SGLD, its defining penalty functions are indeed applicable to gradient-free counterfactual generators. This is an interesting avenue for future research.

Next, common challenges associated with energy-based modelling including sensitivity to scale, training instabilities and sensitivity to hyperparameters also apply to *ECCCo* to some extent. In grid searches for optimal hyperparameters, we have noticed that unless properly regularized, *ECCCo* is sometimes prone to overshoot for the energy constraint.

Finally, while we have used ablation to understand the roles of the different components of *ECCCo*, the scope of this work has prevented us from investigating the role of conformal prediction in this context more thoroughly. We have exclusively relied on split conformal prediction and have used fixed values for the predetermined error rate and other hyperparameters. Future work could benefit from more extensive ablation studies that tune hyperparameters and investigate different approaches to conformal prediction.

4.8. CONCLUSION

This work leverages ideas from energy-based modelling and conformal prediction in the context of counterfactual explanations. We have proposed a new way to generate counterfactuals that are maximally faithful to the black-box model they aim

to explain. Our proposed generator, *ECCCo*, produces plausible counterfactuals iff the black-box model itself has learned realistic explanations for the data, which we have demonstrated through rigorous empirical analysis. This should enable researchers and practitioners to use counterfactuals in order to discern trustworthy models from unreliable ones. While the scope of this work limits its generalizability, we believe that *ECCCo* offers a solid base for future work on faithful counterfactual explanations.

4.9. ACKNOWLEDGEMENTS

Some of the members of TU Delft were partially funded by ICAI AI for Fintech Research, an ING—TU Delft collaboration.

Research reported in this work was partially or completely facilitated by computational resources and support of the DelftBlue ((DHPC) 2022) and the Delft AI Cluster (DAIC: https://doc.daic.tudelft.nl/) at TU Delft. Detailed information about the utilized computing resources can be found in the appendix. The authors would like to thank Azza Ahmed, in particular, for her tremendous help with running Julia jobs on the cluster. The work remains the sole responsibility of the authors.

We would also like to express our gratitude to the group of students who have recently contributed to the development of CounterfactualExplanations.jl (Chapter 2), the Julia package that was used for this analysis: Rauno Arike, Simon Kasdorp, Lauri Keskül, Mariusz Kicior, Vincent Pikand.

All code used for the analysis in this paper can be found here: https://github.com/pat-alt/ECCCo.jl.

5

COUNTERFACTUAL TRAINING: TEACHING MODELS PLAUSIBLE AND ACTIONABLE EXPLANATIONS

We propose a novel training regime termed counterfactual training that leverages counterfactual explanations to increase the explanatory capacity of models. Counterfactual explanations have emerged as a popular post-hoc explanation method for opaque machine learning models: they inform how factual inputs would need to change in order for a model to produce some desired output. To be useful in real-world decision-making systems, counterfactuals should be plausible with respect to the underlying data and actionable with respect to the feature mutability constraints. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. In this work, we instead hold models directly accountable for the desired end goal: counterfactual training employs counterfactuals during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable counterfactual explanations and additionally exhibit improved adversarial robustness.

This chapter will be published in proceedings of the 2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) and will list Patrick Altmeyer, Aleksander Buszydlik, Arie van Deursen and Cynthia C. S. Liem as authors (2026). See Chapter 1.8 for additional publication details.

5.1. INTRODUCTION

Today's prominence of artificial intelligence (AI) has largely been driven by the success of representation learning with high degrees of freedom: instead of relying on features and rules hand-crafted by humans, modern machine learning (ML) models are tasked with learning highly complex representations directly from the data, guided by narrow objectives such as predictive accuracy (Goodfellow, Bengio, and Courville 2016). These models tend to be so complex that humans cannot easily interpret their decision logic.

Counterfactual explanations (CE) have become a key part of the broader explainable AI (XAI) toolkit (Molnar 2022) that can be applied to make sense of this complexity. They prescribe minimal changes for factual inputs that, if implemented, would prompt some fitted model to produce an alternative, more desirable output (Wachter, Mittelstadt, and Russell 2017). This is useful and necessary to not only understand how opaque models make their predictions, but also to provide algorithmic recourse to individuals subjected to them: a retail bank, for example, could use CE to provide meaningful feedback to unsuccessful loan applicants that were rejected based on an opaque automated decision-making (ADM) system (Figure 5.1).

For such feedback to be meaningful, counterfactual explanations need to fulfill certain desiderata (Verma et al. 2022; Karimi et al. 2021)—they should be faithful to the model (Altmeyer, Farmanbar, et al. 2024b), plausible (Joshi et al. 2019), and actionable (Ustun, Spangher, and Liu 2019). Plausibility is typically understood as counterfactuals being *in-domain*: unsuccessful loan applicants that implement the provided recourse should end up with credit profiles that are genuinely similar to that of individuals who have successfully repaid their loans in the past. Actionable explanations further comply with practical constraints: a young, unsuccessful loan applicant cannot increase their age in an instant.

Existing state-of-the-art (SOTA) approaches in the field have largely focused on designing model-agnostic CE methods that identify subsets of counterfactuals, which comply with specific desiderata. This is problematic because the narrow focus on any specific desideratum can adversely affect others: it is possible, for example, to generate plausible counterfactuals for models that are also highly vulnerable to implausible, possibly adversarial counterfactuals (Altmeyer, Farmanbar, et al. 2024b). Indeed, existing approaches generally fail to guarantee that the representations learned by a model are compatible with truly meaningful explanations.

In this work, we propose an approach to bridge this gap, embracing the paradigm that models—as opposed to explanation methods—should be held accountable for explanations that are plausible and actionable. While previous work has shown that at least plausibility can be indirectly achieved through existing techniques aimed at models' generative capacity, generalization and robustness (Altmeyer, Farmanbar, et al. 2024b; Augustin, Meinke, and Hein 2020; Schut et al. 2021), we directly incorporate both plausibility and actionability in the training objective of models to improve their overall explanatory capacity.

Specifically, we introduce **counterfactual training (CT)**: a novel training regime that leverages counterfactual explanations on-the-fly to ensure that differentiable models learn plausible and actionable explanations for the underlying data, while at the same time being more robust to adversarial examples (AE). Figure 5.1 illustrates the outcomes of CT compared to a conventionally trained model. First, in panel (a), faithful and valid counterfactuals end up near the decision boundary forming a clearly distinguishable cluster in the target class (orange). In panel (b), CT is applied to the same underlying linear classifier architecture resulting in much more plausible counterfactuals. In panel (c), the classifier is again trained conventionally and we have introduced a mutability constraint on the *age* feature at test time—counterfactuals are valid but the classifier is roughly equally sensitive to both features. By contrast, the decision boundary in panel (d) has tilted, making the model trained with CT relatively less sensitive to the immutable *age* feature. To achieve these outcomes, CT draws inspiration from the literature on contrastive and robust learning: we contrast faithful CEs with ground-truth data while protecting immutable features, and capitalize on methodological links between CE and AE by penalizing the model's adversarial loss on interim (*nascent*) counterfactuals. To the best of our knowledge, CT represents the first venture in this direction with promising empirical and theoretical results.

The remainder of this manuscript is structured as follows. Section 5.2 presents related work, focusing on the links to contrastive and robust learning. Then follow our two principal contributions. In Section 5.3, we introduce our methodological framework and show theoretically that it can be employed to respect global actionability constraints. In our experiments (Section 5.4), we find that thanks to counterfactual training, (1) the implausibility of CEs decreases by up to 90%; (2) the cost of reaching valid counterfactuals with protected features decreases by 19% on average; and (3) models' adversarial robustness improves across the board. Finally, we discuss open challenges in Section 5.5 and conclude in Section 5.6.

5.2. Related Literature

To make the desiderata for CT more concrete, we follow previous work, tying the explanatory capacity of models to the quality of CEs that can be generated for them (Altmeyer, Farmanbar, et al. 2024b; Augustin, Meinke, and Hein 2020).

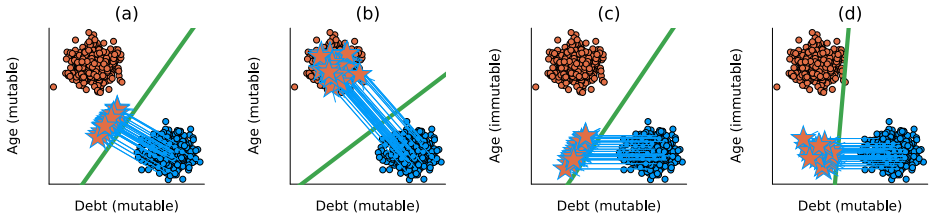


Figure 5.1. Counterfactual explanations (stars) for linear classifiers trained under different regimes on synthetic data: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable. The linear decision boundary is shown in green along with training data colored according to ground-truth labels: $y^- =$ "loan withheld" (blue) and $y^+ =$ "loan provided" (orange). Class and feature annotations (*debt* and *age*) are for illustrative purposes.

5.2.1. EXPLANATORY CAPACITY AND CONTRASTIVE LEARNING

A closely related work shows that model averaging and, in particular, contrastive model objectives can produce models that have a higher explanatory capacity, and hence ones that are more trustworthy (Altmeyer, Farmanbar, et al. 2024b). The authors propose a way to generate counterfactuals that are maximally faithful in that they are consistent with what models have learned about the underlying data. Formally, they rely on tools from energy-based modelling (Teh et al. 2003) to minimize the contrastive divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by a model. Their algorithm, *ECCCo*, yields plausible counterfactual explanations if and only if the underlying model has learned representations that align with them. The authors find that both deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and joint energy-based models (JEMs) (Grathwohl et al. 2020), a form of contrastive learning, do well in this regard.

It helps to look at these findings through the lens of representation learning with high degrees of freedom. Deep ensembles are approximate Bayesian model averages, which are particularly effective when models are underspecified by the available data (Wilson 2020). Averaging across solutions mitigates the risk of overrelying on a single locally optimal representation that corresponds to semantically meaningless explanations. Likewise, it has been shown that generating plausible ("interpretable") CEs is almost trivial for deep ensembles that have undergone adversarial training (Schut et al. 2021). The case for JEMs is even clearer: they optimize a hybrid objective that induces both high predictive performance and strong generative capacity (Grathwohl et al. 2020), resembling the idea of aligning models with plausible explanations. This was an inspiration for CT.

5.2.2. EXPLANATORY CAPACITY AND ROBUST LEARNING

Prior work has shown that counterfactual explanations tend to be more meaningful ("explainable") if the underlying model is more robust to adversarial examples (Augustin, Meinke, and Hein 2020). Once again, we can make intuitive sense of this finding if we look at adversarial training (AT) through the lens of representation learning with high degrees of freedom: highly complex and flexible models may learn representations that make them sensitive to implausible or even adversarial examples (Szegedy et al. 2014). Thus, by inducing models to "unlearn" susceptibility to such examples, adversarial training can effectively remove implausible explanations from the solution space.

This interpretation of the link between explanatory capacity through counterfactuals on the one side, and robustness to adversarial examples on the other is backed by empirical evidence. Firstly, prior work has shown that using counterfactual images during classifier training improves model robustness (Sauer and Geiger 2021). Similarly, related work has shown that counterfactuals represent potentially useful training data in machine learning tasks, especially in supervised settings where inputs may be reasonably mapped to multiple outputs (Abbasnejad et al. 2020). The authors show that augmenting the training data of (image) classifiers can improve generalization performance. Finally, another related work has demonstrated that counterfactual pairs tend to exist in training data (Teney, Abbasnedjad, and Hengel 2020). Hence, the proposed approach aims to identify similar inputs with different annotations and ensure that the gradient of the classifier aligns with the vector between such pairs of inputs using a cosine distance loss function.

CEs have also been used to improve models in the natural language processing domain. A well-known paper in this domain has proposed *Polyjuice* (Wu et al. 2021), a general-purpose CE generator for language models. The authors demonstrate that the augmentation of training data with *Polyjuice* improves robustness in a number of tasks. Related work has introduced the *Counterfactual Adversarial Training* (CAT) framework (Luu and Inoue 2023), which aims to improve generalization and robustness of language models by generating counterfactuals for training samples that are subject to high predictive uncertainty.

There have also been several attempts at formalizing the relationship between counterfactual explanations and adversarial examples. Pointing to clear similarities in how CEs and AEs are generated, prior work makes the case for jointly studying the opaqueness and robustness problems in representation learning (Freiesleben 2022). Formally, the authors show that AEs can be seen as the subset of CEs for which misclassification is achieved (Freiesleben 2022). Similarly, others have shown that CEs and AEs are equivalent under certain conditions (Pawelczyk et al. 2022).

Two other works are closely related to ours in that they use counterfactuals during training with the explicit goal of affecting certain properties of the post-hoc counterfactual explanations. The first closely related work has proposed a way to train models that guarantee recourse to a positive target class with high probability

(Ross, Lakkaraju, and Bastani 2024). The approach builds on adversarial training by explicitly inducing susceptibility to targeted AEs for the positive class. Additionally, the method allows for imposing a set of actionability constraints ex-ante. For example, users can specify that certain features are immutable. A second closely related work has introduced the first end-to-end training pipeline that includes CEs as part of the training procedure (Guo, Nguyen, and Yadav 2023); the *CounterNet* network architecture includes a predictor and a CE generator, where the parameters of the CE generator are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor. In contrast, we impose no restrictions on the artificial neural network architecture at all.

5.3. COUNTERFACTUAL TRAINING

This section introduces the counterfactual training framework, applying ideas from contrastive and robust learning to counterfactual explanations. CT produces models whose learned representations align with plausible explanations that comply with user-defined actionability constraints.

Counterfactual explanations are typically generated by solving variations of the following optimization problem,

$$\min_{\mathbf{x}' \in \mathcal{X}^D} \{\text{yloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}^+) + \lambda \text{reg}(\mathbf{x}')\} \tag{5.1}$$

where $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denotes a classifier, $\mathbf{x}'$ denotes the counterfactual with D features and $\mathbf{y}^+ \in \mathcal{Y}$ denotes some target class. The yloss$(\cdot)$ function quantifies the discrepancy between current model predictions for $\mathbf{x}'$ and the target class (a conventional choice is cross-entropy). Finally, we use reg$(\cdot)$ to denote any form of regularization used to induce certain properties on the counterfactual. The seminal CE paper, (Wachter, Mittelstadt, and Russell 2017), proposes regularizing the distance between counterfactuals and their original factual values to ensure that individuals seeking recourse through CE face minimal costs in terms of feature changes. Different variations of Equation 5.1 have been proposed in the literature to address many desiderata including the ones discussed above (faithfulness, plausibility and actionability). Much like in the seminal work (Wachter, Mittelstadt, and Russell 2017), most of these approaches rely on gradient descent to optimize Equation 5.1, and this holds true for all approaches tested in this work. We introduce them briefly in Section 5.4.1, but refer the reader to the supplementary appendix for details. In the following, we describe how counterfactuals are generated and used in CT.

5.3.1. PROPOSED TRAINING OBJECTIVE

The goal of CT is to improve the explanatory capacity of models by aligning the learned representations with faithful explanations that are plausible and actionable.

For simplicity, we refer to models with high explanatory capacity as **explainable** in this manuscript. We define explainability as follows:

Definition 5.1 (Model Explainability)**.** Let $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denote a supervised classification model that maps from the D-dimensional input space $\mathcal{X}$ to representations $\phi(\mathbf{x};\theta)$ and finally to the K-dimensional output space $\mathcal{Y}$. Let $\mathbf{x}_0'$ denote a factual input and assume that for any given input-output pair $\{\mathbf{x}_0', \mathbf{y}\}_i$ there exists a counterfactual $\mathbf{x}' = \mathbf{x}_0' + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$, where $\arg\max_y \mathbf{y}^+ = y^+$ is the index of the target class.

We say that $\mathbf{M}_\theta$ has an **explanatory capacity** to the extent that faithfully generated, valid counterfactuals are also plausible and actionable. We define these properties as:

- (Faithfulness) $P(\mathbf{x}' \in \mathcal{X}_\theta|\mathbf{y}^+) = 1 - \delta$, where δ is some small value, and $\mathcal{X}_\theta|\mathbf{y}^+$ is the conditional posterior distribution over inputs (adapted from (Altmeyer, Farmanbar, et al. 2024b), Def. 4.1).
- (Plausibility) $P(\mathbf{x}' \in \mathcal{X}|\mathbf{y}^+) = 1 - \delta$, where δ is some small value, and $\mathcal{X}|\mathbf{y}^+$ is the conditional distribution of inputs in the target class (adapted from (Altmeyer, Farmanbar, et al. 2024b), Def. 2.1).
- (Actionability) Perturbations Δ may be subject to some actionability constraints.

Intuitively, plausible counterfactuals are consistent with the data, and faithful counterfactuals are consistent with what the model has learned about the input data. Actionability constraints in Definition 5.1 depend on the context in which $\mathbf{M}_\theta$ is deployed (e.g., specified by end-users or model owners). We consider two types of actionability constraints: on the domain of features and on their mutability. The former naturally arise in automated decision-making systems whenever a feature can only take a specific range of values. For example, *age* is lower bounded by zero and upper bounded by the maximum human lifespan. Specifying such domain constraints can also help address training instabilities commonly associated with energy-based modelling (Grathwohl et al. 2020). The latter arise when a feature cannot be freely modified. Continuing the example, *age* of a person can only increase, but it may even be considered as an immutable feature: waiting many years for an improved outcome is hardly feasible for individuals affected by algorithmic decisions. We choose to only consider domain and mutability constraints for individual features x_d for $d = 1, ..., D$. Of course, this is a simplification since feature values may correlate, e.g., higher *age* may be associated with higher *level of completed education*. We address this challenge in Section 5.5, where we also explain why we restrict this work to classification settings.

Let $\mathbf{x}_t'$ for $t = 0, ..., T$ denote a counterfactual generated through gradient descent over T iterations as originally proposed (Wachter, Mittelstadt, and Russell 2017). CT adopts gradient-based CE search in training to generate on-the-fly model explanations $\mathbf{x}'$ for the training samples. We use the term *nascent* to denote interim

counterfactuals $\mathbf{x}'_{\text{AE}}$ that have not yet converged. As we explain below, these nascent counterfactuals can be stored and repurposed as adversarial examples. Conversely, we consider counterfactuals $\mathbf{x}'_{\text{CE}}$ as *mature* explanations if they have converged within the T iterations by reaching a pre-specified threshold, τ, for the predicted probability of the target class: $\mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau$, where $\mathcal{S}$ is the softmax function.

Formally, we propose the following counterfactual training objective to train explainable (as in Definition 5.1) models,

$$\begin{aligned} &\min_\theta \text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}}\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y^+; \theta) \\ &+\lambda_{\text{adv}}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{\text{AE}}), \mathbf{y}_{\text{AE}}) + \lambda_{\text{reg}}\text{ridge}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta) \end{aligned} \tag{5.2}$$

where yloss$(\cdot)$ is any classification loss that induces discriminative performance (e.g., cross-entropy). The second and third terms are explained in detail in the following subsections. For now, they can be summarized as inducing explainability directly and indirectly by penalizing (1) the contrastive divergence, div$(\cdot)$, between mature counterfactuals $\mathbf{x}'_{\text{CE}}$ and observed samples $\mathbf{x}^+ \in \mathcal{X}^+ = \{\mathbf{x} : y = y^+\}$ in the target class y^+, and (2) the adversarial loss, advloss(.), wrt. nascent counterfactuals $\mathbf{x}'_{\text{AE}}$ and their corresponding labels $\mathbf{y}_{\text{AE}}$. Finally, ridge$(\cdot)$ denotes a Ridge penalty (squared ℓ_2-norm) that regularizes the magnitude of the energy terms involved in the contrastive divergence, div$(\cdot)$, term (Du and Mordatch 2020):

$$\frac{1}{n_{\text{CE}}} \sum_{i=1}^{n_{\text{CE}}} (\mathcal{E}_\theta(\mathbf{x}^+, y^+)^2 + \mathcal{E}_\theta(\mathbf{x}'_{\text{CE}}, y^+)^2) \tag{5.3}$$

The trade-offs between these components are adjusted through penalties λ_{div}, λ_{adv}, and λ_{reg}.

The full counterfactual training regime is sketched out in Algorithm 5.1. During each iteration, we do the following steps. Firstly, we randomly draw a subset of $n_{\text{CE}} \leq n$ factuals $\mathbf{x}'_0$ from $\mathbf{X}$ of size n, for which we uniformly draw a target class y^+ (ensuring that it does not coincide with the class currently predicted for $\mathbf{x}'_0$) and a corresponding training sample from the target class, $\mathbf{x}^+ \sim \mathbf{X}^+ = \{\mathbf{x} \in \mathbf{X} : y = y^+\}$. Secondly, we conduct the counterfactual search by solving (Equation 5.1) through gradient descent. Thirdly, we sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from the training dataset $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ for conventional training and distribute the tuples composed of counterfactuals, their target labels and corresponding training samples, as well as adversarial examples and corresponding labels, $(\mathbf{x}'_{\text{CE}_i}, {y^+}_i, \mathbf{x}'_{\text{AE}_i}, \mathbf{y}_{\text{AE}_i}, {\mathbf{x}^+}_i)_{i=1}^{n_{\text{CE}}}$, across the mini-batches. Finally, we backpropagate through (Equation 5.2).

Algorithm 5.1 Pseudo-Code for Counterfactual Training

Require: Training dataset $\mathcal{D}$, initialize model $\mathbf{M}_\theta$
1: **while** not converged **do**
2: Sample $\mathbf{x}'_0 \sim \mathbf{X}$, $y^+ \sim \mathcal{U}(\mathcal{Y})$ and $\mathbf{x}^+ \sim \mathbf{X}^+$
3: **for** $t = 1$ to T **do**
4: Backpropagate $\nabla_{\mathbf{x}'}$ through equation (5.1)
5: Store $\mathbf{x}'_{\text{CE}}, \mathbf{x}'_{\text{AE}}, \mathbf{y}_{\text{AE}}$
6: **end for**
7: Sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from dataset $\mathcal{D}$
8: Distribute $(\mathbf{x}'_{\text{CE}_i}, {y^+}_i, \mathbf{x}'_{\text{AE}_i}, \mathbf{y}_{\text{AE}_i}, {\mathbf{x}^+}_i)_{i=1}^{n_{\text{CE}}}$
9: **for** each batch **do**
10: Backpropagate ∇_θ through equation (5.2)
11: **end for**
12: **end while**
13: **return** $\mathbf{M}_\theta$

By limiting ourselves to a subset of n_{CE} counterfactuals, we reduce runtimes; this approach has previously been shown to improve efficiency in the context of adversarial training (Kurakin, Goodfellow, and Bengio 2017; Kaufmann et al. 2022). To improve runtimes even more, we choose to first generate counterfactuals and then distribute them across mini-batches to benefit from greater degrees of parallelization during the counterfactual search. Alternatively, it is possible to generate counterfactuals separately for each mini-batch.[1]

5.3.2. Directly Inducing Explainability: Contrastive Divergence

As observed in prior related work (Grathwohl et al. 2020), any classifier can be re-interpreted as a joint energy-based model that learns to discriminate output classes conditional on the observed (training) samples from $p(\mathbf{x})$ and the generated samples from $p_\theta(\mathbf{x})$. The authors show that JEMs can be trained to perform well at both tasks by directly maximizing the joint log-likelihood: $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$, where the first term can be optimized using cross-entropy as in Equation 5.2. To optimize $\log p_\theta(\mathbf{x})$, they minimize the contrastive divergence between the observed samples from $p(\mathbf{x})$ and samples generated from $p_\theta(\mathbf{x})$.

To generate samples, the paper introducing JEMs (Grathwohl et al. 2020) suggests relying on Stochastic Gradient Langevin Dynamics (SGLD) with an uninformative prior for initialization but we depart from this methodology: we propose to leverage counterfactual explainers to generate counterfactuals of observed training samples. Specifically, we have:

[1] During initial prototyping of CT we also tested an implementation that relies on generating counterfactuals and adversarial examples at the batch level with no discernible difference in outcomes, but increased training times.

$$\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y^+; \theta) = \mathcal{E}_\theta(\mathbf{x}^+, y^+) - \mathcal{E}_\theta(\mathbf{x}'_{\text{CE}}, y^+) \tag{5.4}$$

where $\mathcal{E}_\theta(\cdot)$ denotes the energy function defined as $\mathcal{E}_\theta(\mathbf{x}, y^+) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$, with y^+ denoting the index of the randomly drawn target class, $y^+ \sim p(y)$. Conditional on the target class y^+, $\mathbf{x}'_{\text{CE}}$ denotes a mature counterfactual for a randomly sampled factual from a non-target class generated with a gradient-based CE generator for up to T iterations. Intuitively, the gradient of Equation 5.4 decreases the energy of observed training samples (positive samples) while increasing the energy of counterfactuals (negative samples) (Du and Mordatch 2020). As the counterfactuals get more plausible (Definition 5.1) during training, these opposing effects gradually balance each other out (Lippe 2024).

Since the maturity of counterfactuals in terms of a probability threshold is often reached before T, this form of sampling is not only more closely aligned with Definition 5.1., but can also speed up training times compared to SGLD. The departure from SGLD also allows us to tap into the vast repertoire of explainers that have been proposed in the literature to meet different desiderata. For example, many methods support domain and mutability constraints. In principle, any approach for generating CEs is viable, so long as it does not violate the faithfulness condition. Like JEMs (Murphy 2022), counterfactual training can be viewed as a form of contrastive representation learning.

5.3.3. Indirectly Inducing Explainability: Adversarial Robustness

Based on our analysis in Section 5.2, counterfactuals $\mathbf{x}'$ can be repurposed as additional training samples (Balashankar et al. 2023; Luu and Inoue 2023) or adversarial examples (Freiesleben 2022; Pawelczyk et al. 2022). This leaves some flexibility with regards to the choice for the advloss$(\cdot)$ term in Equation 5.2. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

$$\begin{aligned} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{\text{AE}}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t \{t : ||\Delta_t||_\infty < \varepsilon\} \end{aligned} \tag{5.5}$$

Under this choice, we consider nascent counterfactuals $\mathbf{x}'_{\text{AE}}$ as AEs as long as the magnitude of the perturbation at time t (Δ_t) to any single feature is at most ε. The most strongly perturbed counterfactual $\mathbf{x}'_{t_\varepsilon}$ that still satisfies the condition is used as an adversarial example $\mathbf{x}'_{\text{AE}}$. This formalization is closely aligned with seminal work on adversarial machine learning (Szegedy et al. 2014), which defines an adversarial attack as an "imperceptible non-random perturbation". Thus, we work with a different distinction between CE and AE than the one proposed in prior work (Freiesleben 2022), which considers misclassification as the distinguishing feature of adversarial examples. One of the key observations of our work is that we can

leverage CEs during training and get AEs essentially for free to reap the benefits of adversarial training, leading to improved adversarial robustness and plausibility.

5.3.4. ENCODING ACTIONABILITY CONSTRAINTS

Many existing counterfactual explainers support domain and mutability constraints. In fact, both types of constraints can be implemented for any explainer that relies on gradient descent in the feature space for optimization (Altmeyer, Deursen, and Liem 2023a). In this context, domain constraints can be imposed by simply projecting counterfactuals back to the specified domain; if the previous gradient step resulted in updated feature values that were out-of-domain. Similarly, mutability constraints can be enforced by setting partial derivatives to zero to ensure that features are only perturbed in the allowed direction, if at all.

As actionability constraints are binding at test time, we must also impose them when generating $\mathbf{x}'$ during each training iteration to inform model representations. Through their effect on $\mathbf{x}'$, both types of constraints influence model outcomes via Equation 5.4. It is crucial that we avoid penalizing implausibility that arises from mutability constraints. For any mutability-constrained feature d this can be achieved by enforcing $\mathbf{x}^+[d] - \mathbf{x}'[d] := 0$, whenever perturbing $\mathbf{x}'[d]$ in the direction of $\mathbf{x}^+[d]$ would violate mutability constraints defined for d. Specifically, we set $\mathbf{x}^+[d] := \mathbf{x}'[d]$ if:

1. Feature d is strictly immutable in practice.
2. $\mathbf{x}^+[d] > \mathbf{x}'[d]$, but d can only be decreased in practice.
3. $\mathbf{x}^+[d] < \mathbf{x}'[d]$, but d can only be increased in practice.

From a Bayesian perspective, setting $\mathbf{x}^+[d] := \mathbf{x}'[d]$ can be understood as assuming a point mass prior for $p(\mathbf{x}^+)$ with respect to feature d, i.e., we can model this as absolute certainty that the value $\mathbf{x}^+[d]$ remains the same as in the neighbor, $\mathbf{x}'[d]$, but it could be equivalently seen as masking changes to feature d. Intuitively, we can think of this as ignoring implausibility costs of immutable features, which effectively forces the model to instead seek plausibility through the remaining features. This can be expected to produce a classifier with relatively lower sensitivity to immutable features, and the higher relative sensitivity to mutable features should make mutability-constrained recourse less costly (see Section 5.4). Under certain conditions, this result also holds theoretically (for the proof, see the supplementary appendix):

Proposition 5.1 (Protecting Immutable Features)**.** *Let $f_\theta(\mathbf{x}) = \mathcal{S}(\mathbf{M}_\theta(\mathbf{x})) = \mathcal{S}(\Theta\mathbf{x})$ denote a linear classifier with softmax activation $\mathcal{S}$ where $y \in \{1, ..., K\} = \mathcal{K}$, $\mathbf{x} \in \mathbb{R}^D$ and Θ is the matrix of coefficients with $\theta_{k,d} = \Theta[k, d]$ denoting the coefficient on feature d for class k. Assume multivariate Gaussian class densities with a common diagonal covariance matrix $\Sigma_k = \Sigma$ for all $k \in \mathcal{K}$, then protecting an immutable feature from the contrastive divergence penalty will result in lower classifier sensitivity*

to that feature relative to the remaining features, provided that at least one of those is discriminative and mutable.

5.4. EXPERIMENTS

We start by introducing the experimental setup, including performance metrics, datasets, algorithms, and explain our approach to evaluation in Section 5.4.1. Then, we address the research questions (RQ). Two questions relating to the principal goals of counterfactual training are presented in Section 5.4.2:

Research Question 5.1. To what extent does the CT objective in Equation 5.2 induce models to learn plausible explanations?

Research Question 5.2. To what extent does CT result in more favorable algorithmic recourse outcomes in the presence of actionability constraints

Next, in Section 5.4.3 we consider the performance of models trained with CT, focusing on their adversarial robustness but also commenting on the validity of generated CEs.

Research Question 5.3. To what extent does CT influence the adversarial robustness of trained models?

Finally, in Section 5.4.4 we perform an ablation of the CT objective and evaluate its sensitivity to hyperparameters:

Research Question 5.4. How does the CT objective depends on its individual components? (*ablation*)

Research Question 5.5. What are the effects of hyperparameter selection on counterfactual training?

5.4.1. EXPERIMENTAL SETUP

Our focus is the improvement in explainability (Definition 5.1). Thus, we mainly look at the plausibility and cost of faithfully generated counterfactuals at test time, but several other metrics are covered in the supplementary appendix. To measure the cost, we follow the standard proxy of distances (ℓ_1-norm) between factuals and counterfactuals. For plausibility, we assess how similar CEs are to observed samples in the target domain, $\mathbf{X}^+ \subset \mathcal{X}^+$. For the evaluation, we rely on the metric proposed in prior work (Altmeyer, Farmanbar, et al. 2024b) with ℓ_1-norm for distances,

$$\text{IP}(\mathbf{x}', \mathbf{X}^+) = \frac{1}{|\mathbf{X}^+|} \sum_{\mathbf{x} \in \mathbf{X}^+} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{5.6}$$

and introduce a novel divergence-based adaptation,

$$\text{IP}^*(\mathbf{X}', \mathbf{X}^+) = \text{MMD}(\mathbf{X}', \mathbf{X}^+) \tag{5.7}$$

where $\mathbf{X}'$ denotes a collection of counterfactuals and $\text{MMD}(\cdot)$ is the unbiased estimate of the squared population maximum mean discrepancy (Gretton et al. 2012):

$$\begin{aligned} \text{MMD}(\mathbf{X}', \mathbf{X}^+) = & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) \\ & + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(\tilde{x}_i, \tilde{x}_j) \\ & - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, \tilde{x}_j) \end{aligned} \tag{5.8}$$

with a kernel function $k(\cdot, \cdot)$. We use a characteristic Gaussian kernel with a constant length-scale parameter of 0.5, which means that the metric in Equation 5.7 is equal to zero if and only if the two distributions are exactly the same, $\mathbf{X}' = \mathbf{X}^+$.

To assess outcomes with respect to actionability for non-linear models, we look at the costs of (just) valid counterfactuals in terms of their distances from factual starting points with $\tau = 0.5$. While this is an imperfect proxy of sensitivity, we hypothesize that CT can reduce these costs by teaching models to seek plausibility with respect to mutable features, much like we observe in Figure 5.1 in panel (d) compared to (c). We supplement this analysis with estimates using integrated gradients (IG) (Sundararajan, Taly, and Yan 2017). To evaluate predictive performance, we use standard metrics, such as robust accuracy estimated on adversarially perturbed data using the fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2015) and projected gradient descent (PGD) (Madry et al. 2017).

We make use of nine classification datasets common in the CE/AR literature. Four of them are synthetic with two classes and different characteristics: linearly separable Gaussian clusters (*LS*), overlapping clusters (*OL*), concentric circles (*Circ*), and interlocking moons (*Moon*). Next, we have four real-world binary tabular datasets: *Adult* (Census data) (Becker and Kohavi 1996), California housing (*CH*) (Pace and Barry 1997), Default of Credit Card Clients (*Cred*) (Yeh 2016), and Give Me Some Credit (*GMSC*) (Kaggle 2011). Finally, for convenient illustration, we use the 10-class *MNIST* (LeCun 1998).

We run experiments with three gradient-based generators: *Generic* (Wachter, Mittelstadt, and Russell 2017) as a simple baseline; *REVISE* (Joshi et al. 2019) that

aims to generate plausible counterfactuals using a surrogate Variational Autoencoder (VAE); and *ECCCo* (Altmeyer, Farmanbar, et al. 2024b), targeting faithfulness. In all cases, we use standard logit cross-entropy loss for yloss$(\cdot)$ and all generators penalize the distance (ℓ_1-norm) of counterfactuals from their original factual state. *Generic* and *ECCCo* search for counterfactuals directly in the feature space; *REVISE* traverses the latent space of a variational autoencoder (VAE) fitted to the training data, so its outputs depend on the quality of the surrogate model. In addition to the distance penalty, *ECCCo* uses a penalty that regularizes the energy associated with the counterfactual, $\mathbf{x}'$ (Altmeyer, Farmanbar, et al. 2024b). We omit the conformal set size penalty proposed in the original paper, since the authors found that faithfulness primarily depends on the energy penalty, freeing us from one additional hyperparameter.

Our method does not aim to be agnostic to the underlying CE generator and, as explained in Section 5.3.2, the selection of the CE generator can impact the explainability of models. To evaluate the specific value of counterfactual training, we extensively test the method using the three above-mentioned CE generators, which are characterized by varying complexity and desiderata, and we present the complete results in the supplementary appendix. Indeed, we observe that *ECCCo* outclasses the other two generators as the backbone of CT, generally leading to the highest reduction in implausibility. This is not surprising; the goals of *ECCCo* most closely align with the objectives of CT: maximally faithful explanations should also be the most useful for feedback. Conversely, we cannot expect the model to learn much from counterfactual explanations that largely depend on the quality of the surrogate model that is trained for *REVISE*. Similarly, *Generic* is a very simple baseline that optimizes only for minimal changes of features (measured in the original seminal paper (Wachter, Mittelstadt, and Russell 2017) using median absolute deviation).

Thus, while counterfactual training can be used with any gradient-based CE generator to improve the explainability of the resulting model, in Section 5.4.2 we mainly discuss its effectiveness with *ECCCo*, the strongest identified generator, allowing us to optimize the quality of the models. This constitutes our treatment method, but we still present the complete results for all generators in the supplementary appendix.

To assess the effects of CT, we investigate the improvements in performance metrics when using it on top of a weak baseline (BL), a naively (conventionally) trained multilayer perceptron (*MLP*), as the control method. As we hold all other things constant, this is the best way to get a clear picture of the improvement in explainability that can be directly attributed to CT. It is also consistent with the evaluation practices in the related literature (Goodfellow, Shlens, and Szegedy 2015; Ross, Lakkaraju, and Bastani 2024; Teney, Abbasnedjad, and Hengel 2020).

We also note that counterfactual training involves multiple objectives but our principal goal is high explainability as in Definition 5.1, while improved robustness is a welcome byproduct. We neither aim to outperform state-of-the-art approaches

that target any single one of these objectives, nor do we claim that CT can achieve this. Specifically, we do not aim to beat JEMs with respect to their generative capacity, SOTA robust neural networks with respect to (adversarial) robustness, or (quasi-)Bayesian neural networks with respect to uncertainty quantification. As we have already explained in Section 5.2, existing literature has shown that all of these objectives tend to correlate (explaining some of our positive findings), but we situate counterfactual training squarely in the context of (counterfactual) explainability and algorithmic recourse, where it tackles an important shortcoming of existing approaches.

In terms of computing resources, all of our experiments were executed on a high-performance cluster. We have relied on distributed computing across multiple central processing units (CPU); for example, the hyperparameter grid searches were carried out on 34 CPUs with 2GB memory each. Graphical processing units (GPU) were *not* used. All computations were performed in the Julia Programming Language (Bezanson et al. 2017); our code base (algorithms and experimental settings) has been open-sourced on GitHub.[2] We explain more about the hardware, software, and reproducibility considerations in the supplementary appendix.

5.4.2. MAIN RESULTS

Our main results for plausibility and actionability for *MLP* models are summarized in Table 5.1 that presents counterfactual outcomes grouped by dataset along with standard errors averaged across bootstrap samples. Asterisks (*) are used when the bootstrapped 99%-confidence interval of differences in mean outcomes does *not* include zero, so the observed effects are statistically significant at the 0.01 level. As our experimental procedure is (by virtue of the proposed method) relatively complex, we choose to work at this stringent alpha level to demonstrate the high reliability of counterfactual training.

The first two columns (IP and IP*) show the percentage reduction in implausibility for our two metrics when using CT on top of the weak baseline. As an example, consider the first row for *LS* data: the observed positive values indicate that faithful counterfactuals are around 26-51% more plausible for models trained with CT, in line with our observations in panel (b) of Figure 5.1 compared to panel (a).

The third column shows the results for a scenario when mutability constraints are imposed on the selected features. Again, we are comparing CT to the baseline, so reductions in the positive direction imply that valid counterfactuals are "cheaper" (more actionable) when using CT with feature protection. Relating this back to Figure 5.1, the third column represents the reduction in distances traveled by counterfactuals in panel (d) compared to panel (c). In the following paragraphs, we summarize the results for all datasets.

[2] https://github.com/JuliaTrustworthyAI/CounterfactualTraining.jl

Table 5.1. Key evaluation metrics for valid counterfactual along with bootstrapped standard errors for all datasets. **Plausibility** (columns 1-2): percentage reduction in implausibility for **IP** and **IP***, respectively; **Cost** / **Actionability** (column 3): percentage reduction in costs when selected features are protected. Outcomes are aggregated across bootstrap samples (100 rounds) and varying degrees of the energy penalty λ_{egy} used for ECCCo at test time. Asterisks (*) indicate that the bootstrapped 99%-confidence interval of differences in mean outcomes does **not** include zero.

Data	**IP** $(-\%)$	**IP*** $(-\%)$	**Cost** $(-\%)$
LS	$26.26 \pm 0.67^*$	$51.28 \pm 2.01^*$	$16.41 \pm 0.57^*$
Circ	$58.88 \pm 0.37^*$	$93.84 \pm 6.70^*$	$42.99 \pm 0.85^*$
Moon	$19.59 \pm 0.73^*$	8.00 ± 9.44	$5.16 \pm 1.00^*$
OL	-1.93 ± 1.12	-27.70 ± 14.59	$40.86 \pm 2.30^*$
Adult	0.19 ± 1.05	$34.35 \pm 5.61^*$	4.03 ± 4.03
CH	$10.65 \pm 1.47^*$	$63.06 \pm 4.25^*$	$44.23 \pm 1.43^*$
Cred	$10.14 \pm 1.59^*$	$50.35 \pm 12.26^*$	$-18.17 \pm 4.40^*$
GMSC	$10.65 \pm 2.28^*$	$24.75 \pm 4.84^*$	$66.01 \pm 1.41^*$
MNIST	$6.36 \pm 1.70^*$	-70.31 ± 217.60	$-35.11 \pm 6.96^*$
Avg.	15.64	25.29	18.49

PLAUSIBILITY (RQ 5.1)

CT generally produces substantial and statistically significant improvements in plausibility.

Average reductions in IP range from around 6% for *MNIST* to almost 60% for *Circ*. For the real-world tabular datasets they are around 10% for *CH*, *Cred* and *GMSC*; for *Adult* and *OL* we find no significant impact of CT on IP. The former is subject to a large proportion of categorical features, which inhibits the generation of large numbers of valid counterfactuals during training and may therefore explain this finding.

Reductions in IP* are even more substantial and generally statistically significant, although the average degree of uncertainty is higher than for IP: reductions range from around 25% (*GMSC*) to more than 90% (*Circ*). The only negative findings are for *OL* and *MNIST*, but they are insignificant. A qualitative inspection of the counterfactuals in Figure 5.2 suggests recognizable digits for the model trained with CT (bottom row), unlike the baseline (top row).

ACTIONABILITY (RQ 5.2)

CT tends to improve actionability in the presence of immutable features, but this is not guaranteed if the assumptions in Proposition 5.1 are violated.

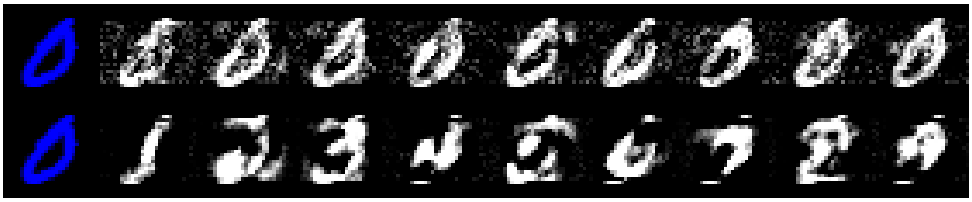


Figure 5.2. *Plausibility*: BL (top row) vs CT using the *ECCCo* generator (bottom row) counterfactuals for a randomly selected factual from class "0" (in blue). CT produces more plausible counterfactuals than BL.

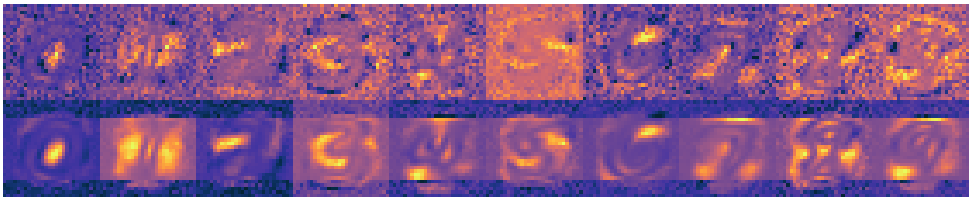


Figure 5.3. *Actionability*: Sample visual explanations (integrated gradients) for all classes in the *MNIST* dataset. Top and bottom rows of images show the results for BL and CT, respectively. Mutability constraints are imposed on the five top and five bottom rows of pixels. CT is less sensitive to protected features.

For synthetic datasets, we always protect the first feature; for all real-world tabular datasets we could identify and protect an *age* variable; for *MNIST*, we protect the five top and five bottom rows of pixels of the full image. Statistically significant reductions in costs overwhelmingly point in the positive direction reaching up to around 66% for *GMSC* data. Only in the case of *Cred* and *MNIST*, average costs increase, most likely because any benefits from protecting features are outweighed by an increase in costs required for greater plausibility. With respect to *MNIST* in particular, the weak baseline is susceptible to cheap adversarial attacks that significantly less costly to achieve that plausible counterfactuals. Finally, the findings for *Adult* are insignificant.

To further empirically evaluate the feature protection mechanism of CT beyond linear models covered in Proposition 5.1, we make use of integrated gradients (IG) (Sundararajan, Taly, and Yan 2017). IG calculates the contribution of each input feature towards a specific prediction by approximating the integral of the model output with respect to its input, using a set of samples that linearly interpolate between a test instance and some baseline instance. This process produces a vector of real numbers, one per input feature, which informs about the contribution of each feature to the prediction. The selection of an appropriate baseline is an important design decision (Sundararajan, Taly, and Yan 2017); to remain consistent in our evaluations, we use a baseline drawn at random from the uniform distribu-

tion $\mathcal{U}(-1, 1)$ for all datasets, which aligns with standard evaluation practices for IG. As the outputs are not bounded (i.e., they are real numbers), we standardize the integrated gradients across features to allow for a meaningful comparison of the results for different models.

Qualitatively, the class-conditional integrated gradients in Figure 5.3 suggest that CT has the expected effect even for non-linear models: the model trained with CT (bottom row) is less sensitive (blue) to the five top and five bottom rows of pixels that were protected. Quantitatively, we observe substantial improvements for seven out of nine datasets, and inconclusive results for the remaining two datasets. Table 5.2 shows the average sensitivity to protected features measured by standardized integrated gradients for CT and BL along with 95% bootstrap confidence intervals: for the synthetic datasets, we observe strong reductions in sensitivity to the protected features for *LS*, *OL* and *OL*, in line with expectations. For the *Moon* dataset, the effect of feature protection is less pronounced but still in the expected direction. We also observe that confidence intervals are in some cases much tighter for models trained with CT: less noisy estimates for integrated gradients likely indicate that the model is more regularized and can be expected to behave more consistently across samples.

For real-world datasets, the sensitivity to the protected *age* variable is reduced by approximately a third for *Adult*, 20% for *CH*, and more than half for protected pixels in *MNIST*, mirroring the qualitative findings in Figure 5.3. In case of *Cred*, CT fully prevents the model from considering *age* as a factor in classification, with sensitivity reduced to zero. Only for *GMSC*, we observe negative impacts of CT, which we believe is due to any or all of the following: a) data assumptions are violated; b) the impact of other components of the CT objective outweighs expected effects of feature protection; or c) the baseline choice applied consistently to all datasets is not appropriate for *GMSC*.

5.4.3. Predictive Performance

Adversarial Robustness (RQ 5.3)

Models trained with CT are much more robust to gradient-based adversarial attacks than conventionally-trained (weak) baselines.

Test accuracies on clean and adversarially perturbed test data are shown in Figure 5.4. The perturbation size, $\varepsilon \in [0, 0.1]$, increases along the horizontal axis, where the case of $\varepsilon = 0$ corresponds to standard test accuracy for non-perturbed data. For synthetic datasets, predictive performance is virtually unaffected by perturbations for all models; those results are therefore omitted from Figure 5.4 in favor of better illustrations for the real-world data.

Focusing on the curves for CT and BL in Figure 5.4 for the moment,[3] we find that standard test accuracy ($\varepsilon = 0$) is largely unaffected by CT, while robustness

[3]The results for AR and CD are discussed in the context of ablation below.

Table 5.2. Median sensitivity to protected features measured by standardized integrated gradients. Square brackets enclose 95% bootstrap confidence intervals.

Dataset	**CT**		**BL**	
LS	0.21	[0.20, 0.22]	30.69	[12.92, 629.20]
Circ	6.96	[4.88, 20.62]	19.20	[6.48, 193.92]
Moons	0.54	[0.41, 0.68]	0.66	[0.53, 0.92]
Over	0.59	[0.38, 0.79]	24.55	[8.31, 466.26]
Adult	0.48	[0.41, 0.52]	0.74	[0.56, 0.91]
CH	0.04	[0.01, 0.06]	0.05	[0.03, 0.09]
Cred	0.00	[0.00, 0.00]	0.20	[0.18, 0.25]
GMSC	0.71	[0.58, 0.85]	0.16	[0.11, 0.23]
MNIST	0.17	[0.16, 0.17]	0.35	[0.33, 0.37]

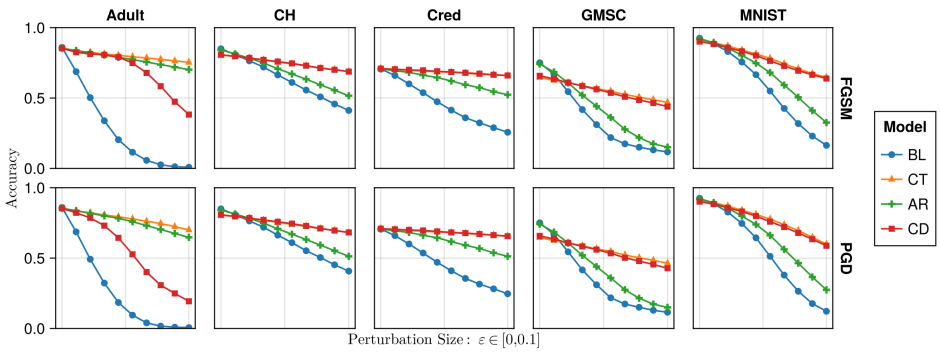


Figure 5.4. Test accuracies on adversarially perturbed data with varying perturbation sizes for the non-synthetic datasets. Different training objectives are distinguished by color and shape: (1) BL—the weak baseline; (2) CT—the full CT objective; (3) AR—a partial CT objective without contrastive divergence; (4) CD—a partial CT objective without adversarial loss. Top and bottom rows show the results for FGSM and PGD (40 steps at step size $\eta = 0.01$), respectively.

against both types of attacks—FGSM (top row) and PGD (bottom row)—is greatly improved: while in some cases robust accuracies for the weak baseline drop to virtually zero (worse than random guessing) for large enough perturbation sizes, accuracies of CT models remain remarkably robust, even though robustness is not

the primary objective of counterfactual training. In the only case where standard accuracy on unperturbed test data is substantially reduced for CT (*GSMC*), we note that robust accuracy decreases particularly fast for the weak baseline as the perturbation size increases. This seems to indicate that the standard accuracy for the weak baseline is inflated by sensitivity to meaningless associations in the data.

We also look at the validity of generated counterfactuals, or the proportion of counterfactuals that attain the target class, as presented in Table 5.3. We find that in many cases CT leads to substantial reductions in average validity, but this effect does not seem to be strongly influenced by the imposed mutability constraints (columns 1-2 vs columns 3-4). This result does not surprise us: by design, CT shrinks the solution space for valid counterfactual explanations, thus making it "harder" (and yet not "more costly") to reach validity compared to the baseline model. As further discussed in the supplementary appendix, this should not be seen as a shortcoming of the method for a number of reasons: validity rates can be increased with longer searches; costs of found solutions still generally decrease, as we observe in our experiments; and achieving high validity does not entail that explanations are practical for the recipients (e.g., valid solutions may still be extremely costly) (Venkatasubramanian and Alfano 2020).

Table 5.3. Average validity of counterfactuals for CT vs BL. First two columns correspond to no mutability constraints imposed on the features; last two columns involve mutability constraints imposed on the specified features.

Data	CT mut.	BL mut.	CT constr.	BL constr.
LS	1.0	1.0	1.0	1.0
Circ	1.0	0.51	0.71	0.48
Moon	1.0	1.0	1.0	0.98
OL	0.86	0.98	0.34	0.56
Adult	0.68	0.99	0.7	0.99
CH	1.0	1.0	1.0	1.0
Cred	0.72	1.0	0.74	1.0
GMSC	0.94	1.0	0.97	1.0
MNIST	1.0	1.0	1.0	1.0
Avg.	0.91	0.94	0.83	0.89

5.4.4. Ablation and Hyperparameter Settings

In this subsection, we use ablation studies to investigate how the different components of the counterfactual training objective in Equation 5.2 affect outcomes. Beyond this, we are also interested in understanding how CT depends on various other hyperparameters. To this end, we present the results from extensive grid searches run across all synthetic datasets.

ABLATION (RQ 5.4)

All components of the CT objective affect outcomes, even independently, but the full objective achieves the most consistent improvements wrt. our goals.

We ablate the effect of both (1) the contrastive divergence component and (2) the adversarial loss included in the full CT objective in Equation 5.2. In the following, we refer to the resulting partial objectives as adversarial robustness (AR) and contrastive divergence (CD), respectively. We note that AR corresponds to a form of adversarial training and the CD objective is similar to that of a joint energy-based model. Therefore, the ablation also serves as a comparison of counterfactual training to stronger baselines, although we emphasize again that we do not seek to outperform SOTA methods in the domains of generative or robust machine learning, focusing CT squarely on models with high explainability and actionability in the context of algorithmic recourse.

Firstly, we find that both components play an important role in shaping final outcomes. Both AR and CD can independently improve the plausibility and adversarial robustness of models.

Concerning plausibility, Figure 5.5 shows the percentage reductions in implausibility for the partial and full objectives compared to the weak baseline. The results for IP and IP* are shown in the top and bottom graphs, respectively, and the datasets are differentiated by color. We find that in the best identified hyperparameter settings, results for the full objective are predominantly affected by the contrastive divergence component, but the inclusion of adversarial loss leads to additional improvements for some datasets (*Adult*, *MNIST*). We penalize contrastive divergence twice as strongly as adversarial loss, which may explain why the former dominates. The outcome for *Adult*, in particular, demonstrates the benefit of including both components: as noted earlier, the large proportion of categorical features in this dataset seems to inhibit the generation of valid counterfactuals, which in turn appears to diminish the effect of the contrastive divergence component.

Looking at AR alone, we find that it produces mixed results for IP, with strong positive results nonetheless dominating overall, reflecting previous findings from the related literature. In particular, for real-world tabular datasets, adversarial robustness seems to substantially benefit plausibility. In these cases, the inclusion of the AR component in the full objective also helps to substantially improve outcomes in relation to the partial CD objective: improvements in plausibility for the *Adult* and *MNIST* datasets are notably higher for full CT. In some cases—most notably *GMSC* and *Cred*—the full CT objective does not outperform the partial objectives, but still achieves the highest levels of adversarial robustness (Figure 5.4).

Zooming in on adversarial robustness, we find that the full CT objective consistently outperforms the partial objectives, which individually yield improvements. Consistent with the existing literature on JEMs (Grathwohl et al. 2020), CD yields substantially more robust models than the weak baseline at varying perturbation sizes

(Figure 5.4). Similarly, AR yields consistent improvements in robustness, as expected. Still, we observe that in cases where either CD or AR show signs of degrading robust accuracy at higher perturbation sizes, the full CT objective maintains robustness. Much like in the context of plausibility, CT benefits from both components, highlighting the effectiveness of our approach to reusing nascent counterfactuals as AEs.

In summary, we find that the full CT objective strikes a balance between both components, thereby leading to the most consistent improvements with respect to plausibility and adversarial robustness.

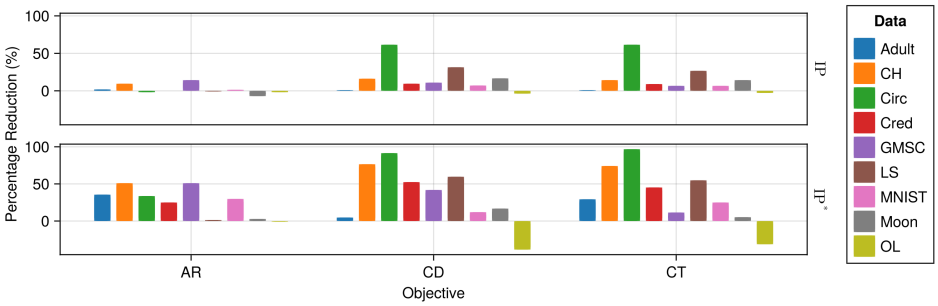


Figure 5.5. Percentage reductions in implausibility for the partial (AR, CD) and full (CT) objectives compared to the weak baseline. The results for IP and IP* are shown in the top and bottom graphs, respectively, and the datasets are differentiated by color.

HYPERPARAMETER SETTINGS (RQ 5.5)

CT is quite sensitive to the choice of a CE generator and its hyperparameters but (1) we observe manageable patterns, and (2) we can usually identify settings that improve either plausibility or actionability, and typically both of them at the same time.

We evaluate the impacts of three types of hyperparameters on CT. In the following, we focus on the highlights and make the full results available in the supplementary appendix.

Firstly, we find that optimal results are generally obtained when using *ECCCo* to generate counterfactuals. Conversely, using a generator that may inhibit faithfulness (*REVISE*), regularly yields smaller improvements in plausibility and is more likely to even increase implausibility. The results of the grid search for *REVISE* also exhibit higher variability than the results for *ECCCo* and *Generic*. As argued above, this finding confirms our intuition that maximally faithful explanations are most suitable for counterfactual training.

Concerning hyperparameters that guide the gradient-based counterfactual search, we find that increasing T, the maximum number of steps, generally yields better outcomes because more CEs can mature. Relatedly, we also find that the effectiveness and stability of CT is positively associated with the total number of counterfactuals generated during each training epoch. The impact of τ, the decision threshold, is more difficult to predict. On "harder" datasets it may be difficult to satisfy high τ for any given sample (i.e., also factuals) and so increasing this threshold does not seem to correlate with better outcomes. In fact, $\tau = 0.5$ generally leads to optimal results as it is associated with high proportions of mature counterfactuals. This is likely because the special case of $\tau = 0.5$ corresponds to equal class probabilities, so a counterfactual is considered mature when the logit for the target class is higher than the logits for all other classes.

Secondly, the strength of the energy regularization, λ_{reg}, is highly impactful and should be set sufficiently high to avoid common problems associated with exploding gradients. The sensitivity with respect to λ_{div} and λ_{adv} is much less evident. While high values of λ_{reg} may increase the variability in outcomes when combined with high values of λ_{div} or λ_{adv}, this effect is not particularly pronounced. These results mirror our observations from the ablation studies and lend further weight to the argument that CT benefits from both components.

Finally, we also observe desired improvements when CT was combined with conventional training and employed only for the final 50% of epochs of the complete training process. Put differently, CT can improve the explainability of models in a post-hoc, fine-tuning manner.

5.5. Discussion

As our results indicate, counterfactual training achieves its objective of producing models that are more explainable. Nonetheless, these advantages come with certain limitations.

Immutable features may have proxies. We propose a method to modify the sensitivity of a model to certain features, and thus increase the actionability of the generated CEs. However, it requires that model owners define the mutability constraints for (all) features considered by the model. Even if all immutable features are protected, there may exist proxies that are theoretically mutable (and hence should not be protected) but preserve enough information about the principals to hinder these protections. Delineating actionability is a major open challenge in the AR literature (see, e.g., (Venkatasubramanian and Alfano 2020)) impacting the capacity of CT to fulfill its intended goal.

Interventions on features may have implications for fairness. Modifying the sensitivity of a model to certain features may also have implications for the fair and equitable treatment of decision subjects. Model owners could misuse this solution by enforcing explanations based on features that are more difficult to modify by

some (group of) decision subjects. For example, consider the *Adult* dataset used in our experiments, where *workclass* or *education* may be more difficult to change for underprivileged groups. When applied irresponsibly, CT could result in an unfairly assigned burden of recourse (Sharma, Henderson, and Ghosh 2020), threatening the equality of opportunity in the system (Bell et al. 2024). Nonetheless, these phenomena are not specific to CT.

Plausibility is costly. As noted before, more plausible counterfactuals are inevitably more costly (Altmeyer, Farmanbar, et al. 2024b). CT improves plausibility and robustness, but this can negatively affect average costs and validity whenever cheap, implausible, and adversarial explanations are removed from the solution space.

CT increases training times. Just like contrastive and robust learning, CT is more resource-intensive than conventional regimes. Three factors mitigate this effect: (1) CT yields itself to parallel execution; (2) it amortizes the cost of CEs for the training samples; and (3) our preliminary findings suggest that it can be used to fine-tune conventionally-trained models.

We also highlight three key directions for future research. Firstly, it is an interesting challenge to extend CT beyond classification settings. Our formulation relies on the distinction between target and non-target classes, requiring the output space to be discrete. Thus, it does not apply to ML tasks where the change in outcome cannot be readily discretized. Classification remains the focus of CE and algorithmic recourse research; other settings have attracted some interest (e.g., regression (Spooner et al. 2021)), but there is little consensus on how to extend the notion of CEs.

Secondly, our analysis covers CE generators with different characteristics, but it is interesting to extend it to more algorithms, including ones that do not rely on computationally costly gradient-based optimization. This should reduce training costs while possibly preserving the benefits of CT.

Finally, we believe that it is possible to considerably improve hyperparameter selection procedures. Our method benefits from the tuning of certain key hyperparameters but we have relied exclusively on grid searches. Future work on CT could benefit from more sophisticated approaches. Notably, CT is iterative, which makes methods such as Bayesian or gradient-based optimization applicable (see, e.g., (Bischl et al. 2023)).

5.6. CONCLUSION

State-of-the-art machine learning models are prone to learning complex representations that cannot be interpreted by humans. Existing work on counterfactual explanations has largely focused on designing tools to generate plausible and actionable explanations for any model. In this work, we instead hold models accountable for delivering such explanations. We introduce counterfactual training: a novel training regime that integrates recent advances in contrastive learning, adversarial robustness, and CE to incentivize highly explainable models. Through theoretical results

and extensive experiments, we demonstrate that CT satisfies this goal while promoting adversarial robustness of models. Explanations generated from CT-based models are both more plausible (compliant with the underlying data-generating process) and more actionable (compliant with user-specified mutability constraints), and thus meaningful to recipients. In turn, our work highlights the value of simultaneously improving models and their explanations.

5.7. ACKNOWLEDGMENTS

Some of the authors were partially funded by ICAI AI for Fintech Research, an ING—TU Delft collaboration. Research reported in this work was partially facilitated by computational resources and support of the DelftBlue high-performance computing cluster at TU Delft ((DHPC) 2022).

6

POSITION: STOP MAKING UNSCIENTIFIC AGI PERFORMANCE CLAIMS

Developments in the field of Artificial Intelligence (AI), and particularly large language models (LLMs), have created a 'perfect storm' for observing 'sparks' of Artificial General Intelligence (AGI) that are spurious. Like simpler models, LLMs distill meaningful representations in their latent embeddings that have been shown to correlate with external variables. Nonetheless, the correlation of such representations has often been linked to human-like intelligence in the latter but not the former. We probe models of varying complexity including random projections, matrix decompositions, deep autoencoders and transformers: all of them successfully distill information that can be used to predict latent or external variables and yet none of them have previously been linked to AGI. We argue and empirically demonstrate that the finding of meaningful patterns in latent spaces of models cannot be seen as evidence in favor of AGI. Additionally, we review literature from the social sciences that shows that humans are prone to seek such patterns and anthropomorphize. We conclude that both the methodological setup and common public image of AI are ideal for the misinterpretation that correlations between model representations and some variables of interest are 'caused' by the model's understanding of underlying 'ground truth' relationships. We, therefore, call for the academic community to exercise extra caution, and to be keenly aware of principles of academic integrity, in interpreting and communicating about AI research outcomes.

This chapter was published in Proceedings of the 41st International Conference on Machine Learning by Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, Cynthia C. S. Liem (2024). See Chapter 1.8 for additional publication details.

6.1. INTRODUCTION

In 1942, when anti-intellectualism was rising and the integrity of science was under attack, Robert K. Merton formulated four 'institutional imperatives' as comprising the ethos of modern science: *universalism*, meaning that the acceptance or rejection of claims entering the lists of science should not depend on personal or social attributes of the person bringing in these claims; *"communism"* [sic], meaning that there should be common ownership of scientific findings and one should communicate findings, rather than keeping them secret; *disinterestedness*, meaning that scientific integrity is upheld by not having self-interested motivations, and *organized skepticism*, meaning that judgment on the scientific contribution should be suspended until detached scrutiny is performed, according to institutionally accepted criteria (Merton et al. 1942). While the Mertonian norms may not formally be known to academics today, they still are implicitly being subscribed to in many ways in which academia has organized academic scrutiny; e.g., through the adoption of double-blind peer reviewing, and in motivations behind open science reforms.

At the same time, in the way in which academic research is disseminated in the AI and machine learning fields today, major shifts are happening. Where these research fields have actively adopted early sharing of preprints and code, the volume of publishable work has exploded to a degree that one cannot reasonably keep up with broad state-of-the-art, and social media influencers start playing a role in article discovery and citeability (Weissburg et al. 2024). Furthermore, because of major commercial stakes with regard to AI and machine learning technology, and e.g. following the enthusiastic societal uptake of products employing LLMs, such as ChatGPT, the pressure to beat competitors as fast as possible is only increasing, and strong eagerness can be observed in many domains to 'do something with AI' in order to innovate and remain current.

Where AI used to be a computational modeling tool to better understand human cognition (Rooij et al. 2023), the recent interest in AI and LLMs has been turning into one in which AI is seen as a tool that can mimic, surpass and potentially replace human intelligence. In this, the achievement of Artificial General Intelligence (AGI) has become a grand challenge, and in some cases, an explicit business goal. The definition of AGI itself is not as clear-cut or consistent; loosely, it is a phenomenon contrasting with 'narrow AI' systems, that were trained for specific tasks (Goertzel 2014). In practice, to demonstrate that the achievement of AGI may be getting closer, researchers have sought to show that AI models generalize

to different (and possibly unseen) tasks, with little human intervention, or show performance considered 'surprising' to humans.

For example, Google DeepMind claimed their AlphaGeometry model (Trinh et al. 2024) reached a 'milestone' towards AGI. This model has the ability to solve complex geometry problems, allegedly without the need for human demonstrations during training. However, work such as this had been initially introduced in the 1950s (Zenil 2024): without the use of an LLM, logical inference systems proved 100% accurate in proving all the theorems of Euclidean Geometry, due to geometry being an axiomatically closed system. Therefore, while DeepMind created a powerfully fast geometry-solving machine, it is still far from AGI.

Generally, in the popularity of ChatGPT and the integration of generative AI in productivity tools (e.g. through Microsoft's Copilot integrations in GitHub and Office applications), one also can wonder whether the promise of AI is more in computationally achieving general intelligence, or rather in the engineering of general-purpose tools[1]. Regardless, stakes and interests are high, e.g. with ChatGPT clearing nearly $1 billion in months of its release[2].

When combining massive financial incentives with the presence of a challenging and difficult-to-understand technology, that aims towards human-like problem-solving and communication abilities, a situation arises that is fertile for the misinterpretation of spurious cues as hints towards AGI, or other qualities like sentience[3] and consciousness. AI technology only becomes more difficult to understand as academic publishing in the space largely favors performance, generalization, quantitative evidence, efficiency, building on past work, and novelty (Birhane et al. 2022). As such, works that make it into top-tier venues tend to propose heavier and more complicated technical takes on tasks that (in the push towards generalizability) get more vague, while the scaling-up of data makes traceability of possible memorization harder. In a submission-overloaded reality, researchers may further get incentivized to oversell and overstate achievement claims. At the same time, while currently popular in literature, inherent complexity and opaqueness in technical solutions may fundamentally be unwise to pursue in high-stakes applications (Rudin 2019).

Noticing these trends, we as the authors of this article are concerned. We feel that the current culture of racing toward Big Outcome Statements in industry and academic publishing too much disincentivizes efforts toward more thorough and nuanced actual problem understanding. At the same time, as the outside world is so eager to adopt AI technology, (too) strong claims make for good sales pitches, but a question is whether there is indeed sufficient evidence for these claims. With successful AGI outcomes needing to look human-like, this also directly plays into risks of anthropomorphizing (the attribution of human-like qualities to non-human objects)

[1] A Swiss army knife is an effective general-purpose tool, without people wondering whether it exhibits intelligence.

[2] https://www.bloomberg.com/news/articles/2023-08-30/openai-nears-1-billion-of-annual-sales-as-chatgpt-takes-off

[3] https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

and confirmation bias (the seeking-out and/or biased interpretation of evidence in support of one's beliefs). In other words, it is very tempting to claim surprising human-like achievements of AI, and as humans, we are very prone to genuinely believing this. **We therefore urge our fellow researchers to stop making unscientific AGI performance claims**.

To strengthen our argument, in this paper, we first present related work in Section 6.2. We then consider a recently viral work (Gurnee and Tegmark 2023a) in which claims about the learning of world models by LLMs were made. In Section 6.3, we present several experiments that may invite similar claims on models yielding more intelligent outcomes than would have been expected—while at the same time indicating how we feel these claims should *not* be made. Furthermore, we present a review of social science findings in Section 6.4 that underline how prone humans are to being enticed by patterns that are not really there. Combining this with the way in which media portrayal of AI has tended towards science-fiction imagery of mankind-threatening robots, we argue that the current AI culture is a perfect storm for making and believing inflated claims, and call upon our fellow academics to be extra mindful and scrutinous about this. Finally, in Section 6.5, we propose specific structural and cultural changes to improve the current situation. Section 6.6 concludes.

6.2. RELATED WORK

In this work, we question the practice of using outcomes from mechanistic interpretability to support AGI claims. This is not to be seen as criticism toward the underlying methodologies in isolation, but rather in the context of current publishing practices that we repeatedly challenge throughout this work. Many closely related works are free of any grandiose conclusions and instead highlight the benefits of mechanistic interpretability that we also highlight here (Nanda, Lee, and Wattenberg 2023; Gurnee et al. 2023; Li et al. 2022).

Another related subfield investigates the capacity of LLMs to reason causally. Here, too, there is an opportunity to over-interpret the finding of causal information as causal understanding. Recent work has shown, for example, that LLMs can indeed correctly predict causal relationships and this may have practical use cases (Kıcıman et al. 2023). But despite the potential utility, the authors also demonstrate that this capacity can be partially explained by memorization, rather than an actual understanding of causal relationships. Similarly, Zečević et al. (2023) provide evidence indicating that current LLMs "may talk causality but are not causal".

Two other recent works are related to this work and align well with the position we present here. Schaeffer, Miranda, and Koyejo (2024) demonstrate that the apparent emergent abilities of large language models may be driven by a choice of evaluation metrics, rather than some fundamental property that is intrinsic to this family of models. Their work highlights the need for rigorous testing and benchmarking of

LLMs, which we also point to in this work, albeit in a slightly different methodological context. Kloft et al. (2024) provide experimental evidence demonstrating that people have heightened expectations and a biased, positive view of AI. The authors run a user study of human-AI interaction, in which participants performed better at a given task when they (wrongly) thought they were aided by a positively described AI. This placebo effect was found to be robust to negative descriptions of AI. Positive bias towards AI may exacerbate other factors that drive people to make unscientific claims about the current state of AI, which we discuss in Section 6.4.

6.3. Surprising Patterns in Latent Spaces?

In 2023, a research article went viral on the X[4] platform (Gurnee and Tegmark 2023b). Through linear probing experiments, the claim was made that LLMs learned literal maps of the world. As such, they were considered to be more than 'stochastic parrots' (Bender et al. 2021) that can only correlate and mimic existing patterns from data, but not truly understand it. While the manuscript immediately received public criticism (Marcus 2023), and the revised, current version is more careful with regard to its claims (Gurnee and Tegmark 2023a), reactions on X seemed to largely exhibit excitement and surprise at the authors' findings. However, in this section, through various simple examples, we make the point that observing patterns in latent spaces should not be a surprising revelation. After starting with a playful example of how easy it is to 'observe' a world model, we build up a larger example focusing on key economic indicators and central bank communications.

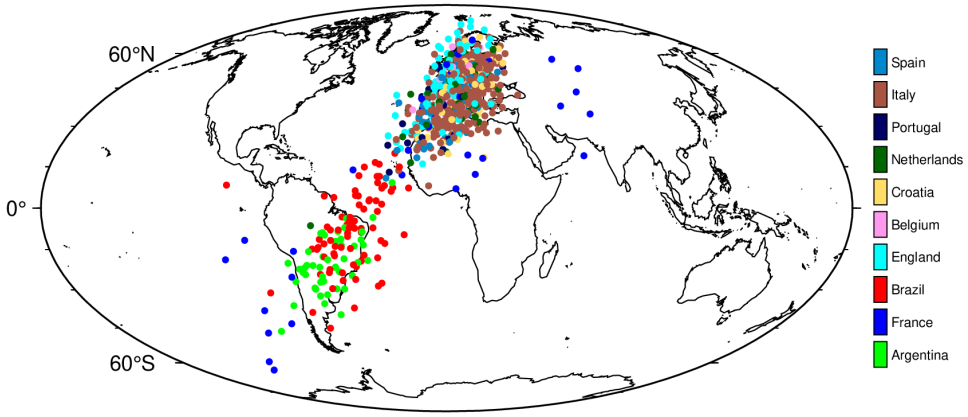


Figure 6.1. Predicted coordinate values (out-of-sample) from a linear probe on final-layer activations of an untrained neural network.

[4]https://twitter.com/wesg52/status/1709551516577902782?s=20

6.3.1. Are neural networks born with world models?

Gurnee and Tegmark (2023a) extract and visualize the alleged geographical world model by training linear regression probes on internal activations in LLMs (including Llama-2) for the names of places, to predict geographical coordinates associated with these places. Now, the Llama-2 model has ingested huge amounts of publicly available data from the internet, including Wikipedia dumps from the June-August 2022 period (Touvron et al. 2023). It is therefore highly likely that the training data contains geographical coordinates, either directly or indirectly. At the very least, we should expect that the model has seen features during training that are highly correlated with geographical coordinates. The model itself is essentially a very large latent space to which all features are randomly projected in the very first instance before being passed through a series of layers which are gradually trained for downstream tasks.

In our first example, we simulate this scenario, stopping short of training the model. In particular, we take the world_place.csv that was used in Gurnee and Tegmark (2023a), which maps locations/areas to their latitude and longitude. For each place, it also indicates the corresponding country. From this, we take the subset that contains countries that are currently part of the top 10 FIFA world ranking, and assign the current rank to each country (i.e., Argentina gets 1, France gets 2, ...). To ensure that the training data only involves a noisy version of the coordinates, we transform the longitude and latitude data as follows: $\rho \cdot \text{coord} + (1-\rho) \cdot \epsilon$ where $\rho = 0.5$ and $\epsilon \sim \mathcal{N}(0, 5)$.

Next, we encode all features except the FIFA world rank indicator as continuous variables: $X^{(n \times m)}$ where n is the number of samples and m is the number of resulting features. Additionally, we add a large number of random features to X to simulate the fact that not all features ingested by Llama-2 are necessarily correlated with geographical coordinates. Let d denote the final number of features, i.e. $d = m + k$ where k is the number of random features.

We then initialize a small neural network, considered a *projector*, mapping from X to a single hidden layer with $h < d$ hidden units and sigmoid activation, and from there, to a lower-dimensional output space. Without performing any training on the *projector*, we simply compute a forward pass of X and retrieve activations $\mathbf{Z}^{(n \times h)}$. Next, we perform the linear probe on a subset of $\mathbf{Z}$ through Ridge regression: $\mathbf{W} = (\mathbf{Z}'_{\text{train}}\mathbf{Z}_{\text{train}} + \lambda\mathbf{I})(\mathbf{Z}'_{\text{train}}\mathbf{coord})^{-1}$, where $\mathbf{coord}$ is the $(n \times 2)$ matrix containing the longitude and latitude for each sample. A hold-out set is reserved for testing, on which we compute predicted coordinates for each sample as $\widehat{\mathbf{coord}} = \mathbf{Z}_{\text{test}}\mathbf{W}$ and plot these on a world map (Figure 6.1).

While the fit certainly is not perfect, the results do indicate that the random projection contains representations that are useful for the task at hand. Thus, this simple example illustrates that meaningful target representations should be recoverable from a sufficiently large latent space, given the projection of a small number of highly correlated features. Similarly, Alain and Bengio (2016) observe that even

before training a convolutional neural network on MNIST data, the layer-wise activations can already be used to perform binary classification. In fact, it is well-known that random projections can be used for prediction tasks (Dasgupta 2013).

This first experiment—and indeed the practice of probing LLMs that have seen vast amounts of data—can be seen as a form of inverse problem and common caveats such as non-uniqueness and instability apply (Haltmeier and Nguyen 2023). Regularization can help mitigate these caveats (OM 2001), but we confess that we did not carefully consider the parameter choice for λ, nor has this been carefully studied in the related literature to the best of our knowledge.

6.3.2. PCA AS A YIELD CURVE INTERPRETER

We now move to a concrete application domain: Economics. Here, the yield curve, plotting the yields of bonds against their maturities, is a popular tool for investors and economists to gauge the health of the economy. The yield curve's slope is often used as a predictor of future economic activity: a steep yield curve is associated with a growing economy, while a flat or inverted yield curve is associated with a contracting economy. To leverage this information in downstream modelling tasks, economists regularly use PCA to extract a low-dimensional projection of the yield curve that captures relevant variation in the data (e.g. Berardi and Plazzi (2022), Kumar (2022) and Crump and Gospodinov (n.d.)).

To understand the nature of this low-dimensional projection, we collect daily Treasury par yield curve rates at all available maturities from the US Department of the Treasury. Computing principal components involves decomposing the matrix of all yields $\mathbf{r}$ into a product of its singular vectors and values: $\mathbf{r} = \mathbf{U}\Sigma\mathbf{V}'$. Let us simply refer to $\mathbf{U}$, Σ and $\mathbf{V}'$ as latent embeddings of the yield curve.

The upper panel in Figure 6.2 shows the first two principal components of the yield curves of US Treasury bonds over time. Vertical stalks indicate key dates related to the Global Financial Crisis (GFC). During its onset, on 27 February 2007, financial markets were in turmoil following a warning from the Federal Reserve (Fed) that the US economy was at risk of a recession. The Fed later reacted to mounting economic pressures by gradually reducing short-term interest rates to unprecedented lows. Consequently, the average level of yields decreased and the curve steepened. In Figure 6.2, we can observe that the first two principal components appear to capture this level shift and steepening, respectively. In fact, they are strongly positively correlated with the actual observed first two moments of the yield curve (lower panel of Figure 6.2).

Again, it should not be surprising that these latent embeddings are meaningful: by construction, principal components are orthogonal linear combinations of the data itself, each of which explains most of the residual variance after controlling for the effect of all previous components.

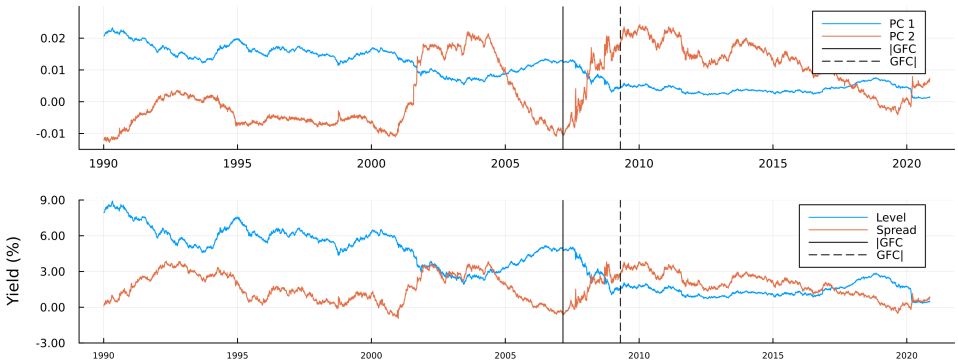


Figure 6.2. Top chart: The first two principal components of US Treasury yields over time at daily frequency. Bottom chart: Observed average level and 10yr-3mo spread of the yield curve. Vertical stalks roughly indicate the onset (|GFC) and the beginning of the aftermath (GFC|) of the Global Financial Crisis.

6

6.3.3. LLMS FOR ECONOMIC SENTIMENT PREDICTION

So far, we considered simple linear data transformations. One might argue that this does not really involve latent embeddings in the way they are typically thought of in the context of deep learning. In the appendix, we present an additional experiment in which we more explicitly seek neural network-based representations that will be useful for downstream tasks. Here, we continue with an example in which LLMs may be used for economic sentiment prediction.

Closely following the approach in Gurnee and Tegmark (2023a), we apply it to the novel *Trillion Dollar Words* (Shah, Paturi, and Chava 2023) financial dataset, containing a curated selection of sentences formulated and communicated to the public by the Fed through speeches, meeting minutes and press conferences. (Shah, Paturi, and Chava 2023) use this dataset to train a set of LLMs and rule-based models to classify sentences as either 'dovish', 'hawkish' or 'neutral'. In the context of central banking, 'hawkishness' is typically associated with tight monetary policy: in other words, a 'hawkish' stance on policy favors high interest rates to limit the supply of money and thereby control inflation. The authors first manually annotate a subsample of the available data and then fine-tune various models for the classification task. Their model of choice, *FOMC-RoBERTa* (a fine-tuned version of RoBERTa (Liu et al. 2019)), achieves an F_1 score of around > 0.7 on the test data. To illustrate the potential usefulness of the learned classifier, they use predicted labels for the entire dataset to compute an ad-hoc, count-based measure of 'hawkishness'. This measure is shown to correlate with key economic indicators in the expected direction: when inflationary pressures rise, the measured level of 'hawkishness' increases, as central bankers react by raising interest rates to bring inflation back to

target.

6.3.3.1. LINEAR PROBES

We now use linear probes to assess if the fine-tuned model has learned associative patterns between central bank communications and key economic indicators. Therefore, we further pre-process the data provided by Shah, Paturi, and Chava (2023) and use their proposed model to compute activations of the hidden state, on the first entity token for each layer. We have made these available and easily accessible through a small Julia package: TrillionDollarWords.jl.

For each layer, we compute linear probes through Ridge regression on two inflation indicators (the Consumer Price Index (CPI) and the Producer Price Index (PPI)) and US Treasury yields at different levels of maturity. To allow comparison with Shah, Paturi, and Chava (2023), we let yields enter the regressions in levels. To measure price inflation we use percentage changes proxied by log differences. To mitigate issues related to over-parameterization, we follow the recommendation in Alain and Bengio (2016) to first reduce the dimensionality of the computed activations. In particular, we restrict our linear probes to the first 128 principal components of the embeddings of each layer. To account for stochasticity, we use an expanding window scheme with 5 folds for each indicator and layer. To avoid look-ahead bias, PCA is always computed on the sub-samples used for training the probe.

Figure 6.3 shows the out-of-sample root mean squared error (RMSE) for the linear probe, plotted against *FOMC-RoBERTa*'s n-th layer. The values correspond to averages across cross-validation folds. Consistent with related work (Alain and Bengio 2016; Gurnee and Tegmark 2023a), we observe that model performance tends to be higher for layers near the end of the transformer model. Curiously, for yields at longer maturities, we find that performance eventually deteriorates for the very final layers. We do not observe this for the training data, so we attribute this to overfitting.

It should be noted that performance improvements are generally of small magnitude. Still, the overall qualitative findings are in line with expectations. Similarly, we also observe that these layers tend to produce predictions that are more positively correlated with the outcome of interest and achieve higher mean directional accuracy (MDA). Upon visual inspection of the predicted values, we conclude the primary source of prediction errors is low overall sensitivity, meaning that the magnitude of predictions is generally too small.

To better assess the predictive power of our probes, we compare their predictions to those made by simple autoregressive models. For each layer, indicator and cross-validation fold, we first determine the optimal lag length based on the training data using the Bayes Information Criterium with a maximal lag length of 10. These are not state-of-the-art forecasting models, but they serve as a reasonable baseline. For most indicators, probe predictions outperform the baseline in terms of average

performance measures. After accounting for variation across folds, however, we generally conclude that the probes neither significantly outperform nor underperform. Detailed results, in which we also perform more explicit statistical testing, can be found in the appendix.

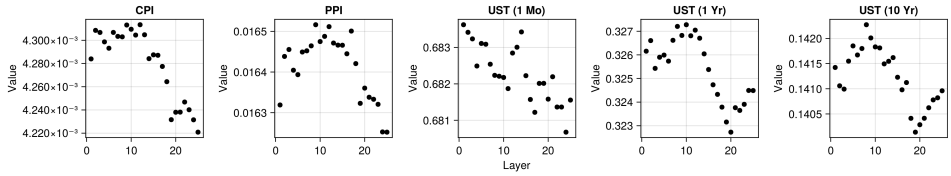


Figure 6.3. Out-of-sample root mean squared error (RMSE) for the linear probe plotted against *FOMC-RoBERTa*'s n-th layer for different indicators. The values correspond to averages computed across cross-validation folds, where we have used an expanding window approach to split the time series. As expected, model performance tends to be higher (average prediction errors are lower) for layers near the end of the transformer model.

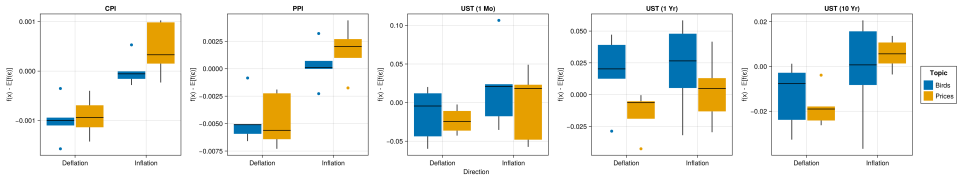


Figure 6.4. Probe predictions for sentences about inflation of prices (IP), deflation of prices (DP), inflation of birds (IB) and deflation of birds (DB). The vertical axis shows predicted inflation levels subtracted by the average predicted value of the probe for random noise.

6.3.3.2. Sparks of Economic Understanding?

Even though *FOMC-RoBERTa*, which is substantially smaller than the models tested in Gurnee and Tegmark (2023a), was not explicitly trained to uncover associations between central bank communications and the level of consumer prices, it appears that the model has distilled representations that can be used to predict inflation (although they certainly will not win any forecasting competitions). So, have we uncovered further evidence that LLMs "aren't mere stochastic parrots"? Has *FOMC-RoBERTa* developed an intrinsic 'understanding' of the economy just by 'reading' central bank communications? Thus, can economists readily adopt FOMC-RoBERTa as a domain-relevant tool?

We are having a very hard time believing that the answer to either of these questions is 'yes'. To argue our case, we will now produce a counter-example demonstrating

that, if anything, these findings are very much in line with the parrot metaphor. The counter-example is based on the following premise: if the results from the linear probe truly were indicative of some intrinsic 'understanding' of the economy, then the probe should not be sensitive to random sentences that are most definitely not related to consumer prices.

To test this, we select the best-performing probe trained on the final-layer activations for each indicator. We then make up sentences that fall into one of these four categories: *Inflation/Prices* (IP)—sentences about price inflation, *Deflation/Prices* (DP)—sentences about price deflation, *Inflation/Birds* (IB)—sentences about inflation in the number of birds and *Deflation/Birds* (DB)—sentences about deflation in the number of birds. A sensible sentence for category DP, for example, could be: "It is essential to bring inflation back to target to avoid drifting into deflation territory.". Analogically, we could construct the following sentence for the DB category: "It is essential to bring the numbers of doves back to target to avoid drifting into dovelation territory.". While domain knowledge suggests that the former is related to actual inflation outcomes, the latter is, of course, completely independent of the level of consumer prices. Detailed information about the made-up sentences can be found in the appendix.

In light of the encouraging results in Figure 6.3, we should expect the probe to predict higher levels of inflation for activations for sentences in the IP category, than for sentences in the DP category. If this was indicative of true intrinsic 'understanding' as opposed to memorization, we would not expect to see any significant difference in predicted inflation levels for sentences about birds, independent of whether their numbers are increasing. More specifically, we would not expect the probe to predict values for sentences about birds that are substantially different from the values it can be expected to predict for actual white noise.

To get to this last point, we also generate many probe predictions for samples of noise. Let $f : \mathcal{A}^k \mapsto \mathcal{Y}$ denote the linear probe that maps from the k-dimensional space spanned by k first principal components of the final-layer activations to the output variable of interest (CPI growth in this case). Then we sample $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^{(k\times k)})$ for $i \in [1, 1000]$ and compute the sample average. We repeat this process 10000 times and compute the median-of-means to get an estimate for $\mathbb{E}[f(\varepsilon)] = \mathbb{E}[y|\varepsilon]$, that is the predicted value of the probe conditional on random noise.

Figure 6.4 shows the results of this small test: it shows predicted inflation levels subtracted by $\mathbb{E}[f(\varepsilon)]$. The median linear probe predictions for sentences about inflation and deflation are indeed substantially higher and lower, respectively than for random noise. Unfortunately, the same is true for sentences about the inflation and deflation in the number of birds, albeit to a somewhat lower degree. This finding holds for both inflation indicators and to a lesser degree also for yields at different maturities, at least qualitatively.

6.4. HUMAN PRONENESS TO OVER-INTERPRETATION

Linear probes and related tools from mechanistic interpretability were proposed in the context of monitoring models and diagnosing potential problems (Alain and Bengio 2016). Favorable outcomes from probes merely indicate that the model "has learned information relevant for the property [of interest]" (Belinkov 2021). Our examples demonstrate that this is achievable even for small models, while these have certainly not developed intrinsic"understanding" of the world. Thus, we argue that more conservative and rigorous tests for emerging capabilities of AI model are needed.

Generally, humans are prone to seek patterns everywhere. Meaningful patterns have proven useful in helping us make sense of our past, navigate our present and predict the future. Although this tendency to perceive patterns likely leads to evolutionary benefits even when the perceived patterns are false (Foster and Kokko 2009), psychology has revealed a host of situations in which the ability to perceive patterns severely misfires, leading to irrational beliefs in the power of superstitions (Foster and Kokko 2009), conspiracy theories (Van Prooijen, Douglas, and De Inocencio 2018), the paranormal (Müller and Hartmann 2023), gambler's fallacies (Ladouceur, Paquet, and Dubé 1996) and 'pseudo-profound bullshit' (Walker et al. 2019).

We argue herein that AI research and development is a perfect storm that encourages our human biases to perceive spurious sparks of general intelligence in AI systems. When an AI system extracts patterns in the corpus not originally (thought to be) perceived during training, we can easily be misled to perceive and interpret this as the AI system having greater cognitive capabilities. We further elaborate on this by highlighting the risks of finding spurious patterns, and reviewing social science knowledge on the tendency of humans to anthropomorphize and have cognitive bias.

6.4.1. SPURIOUS RELATIONSHIPS

In statistics, misleading patterns are often referred to as spurious relationships: associations, often quantitatively assessed, between two or more variables that are not causally related to each other. Although the formal definition of spuriousness varies somewhat (Haig 2003), it distinctly implies that the observation of correlations does not necessarily imply causation. Quantitative data often show non-causal associations (as humorously demonstrated on the Spurious Correlations website), and as adept as humans are at recognizing patterns, we typically have a much harder time discerning spurious relationships from causal ones.

A major contributor is that humans struggle to tell the difference between random and non-random sequences (Falk and Konold 1997), and to generate sequences that appear random (Ladouceur, Paquet, and Dubé 1996). A common issue is a lack of expectation that randomness that hints towards a causal relationship, such as correlations, will still appear at random. This leads even those trained in statistics

and probability to perceive illusory correlations, correlations of inflated magnitude (see Nickerson (1998)), or causal relationships in data that is randomly generated (Zgraggen et al. 2018).

6.4.2. ANTHROPOMORPHISM

Research on anthropomorphism has repeatedly shown the human tendency to attribute human-like characteristics to non-human agents and/or objects. These might include the weather and other natural forces, pets and other animals, gadgets and other pieces of technology (Epley, Waytz, and Cacioppo 2007). Formally studied as early as 1944, Heider and Simmel (1944) observed that humans can correctly interpret a narrative whose characters are abstract 2D shapes, but also that humans interpreted random movements of these shapes as having a human-like narrative. Relevant to AI and the degree to which it resembles AGI, anthropomorphizing may occur independently of whether such judgments are accurate, and as a matter of degree: at the weaker end, one may employ anthropomorphism as a metaphorical way of thinking or explaining, and at the stronger end one may attribute human emotions, cognition, and intelligence to AI systems. As Epley, Waytz, and Cacioppo (2007) note, literature has shown that even weak metaphorical anthropomorphism may affect how humans behave towards non-human agents.

Modern anthropomorphism theory suggests there are three key components, one of which is a cognitive feature, and two of which are motivations. The first involves the easy availability of our experiences as heuristics that can be used to explain external phenomena: "...knowledge about humans in general, or self-knowledge more specifically, functions as the known and often readily accessible base for induction about the properties of unknown agents" (p.866 in Epley, Waytz, and Cacioppo (2007)). Thus, our experience as humans is an always-readily-available template to interpret the world, including non-human agent behaviors. This may be more so when the behaviors of that agent are made to resemble humans, which can be a benefit to the second key component of the theory: a motivational state to anthropomorphize among individuals experiencing loneliness, social isolation, or otherwise seeking social connection (Epley, Waytz, and Cacioppo 2007; Waytz, Epley, and Cacioppo 2010).

The third component is the motivation as a human to be competent (effectance motivation). This is most relevant to this discussion, as it describes the need to effectively interact with our environments, including the technologies of the day (Epley, Waytz, and Cacioppo 2007). When confronted with an opaque technology, a person may interpret its behaviors using the most readily available template at hand, namely their personal human experience, in order to facilitate learning (Epley, Waytz, and Cacioppo 2007; Waytz, Epley, and Cacioppo 2010). Perceiving human characteristics, motivations, emotions, and cognitive processes from one's own experiences in a technology such as an AI chatbot, allows for a ready template of comparison at the very least, and possibly an increase in ability to make sense of,

and even predict, the agent's behaviors. This may include being placed in a position to master a certain technology, whether by incentives to learn, or fear of poor outcomes should one not manage to learn.

These pressures extend to AI experts, as well as laypersons. In both scholarly and commercial fields, AI experts face considerable pressure to demonstrate competence in their work. Citation metrics and scholarly publications remain the primary metric for tenure and promotion (Alperin et al. 2019), and the number of publications in the AI field has boomed as evidenced by overall (preprint and peer-reviewed) scholarly publications[5] (Maslej et al. 2023). The adoption of techniques underlying technologies with the AI label, i.e. machine learning, has spread to fields beyond Computer Science, e.g. Astronomy, Physics, Medicine and Psychology[6]. Outside of academia, the number of jobs requiring AI expertise increases rapidly, with demand for 'Machine Learning' skills clusters having increased over 500% from 2010 to 2020 (Maslej et al. 2023). Thus, according to theory, the pressure to demonstrate AI-competence is fertile ground for anthropomorphism to occur.

6.4.3. CONFIRMATION BIAS

Confirmation bias is generally defined as favoring interpretations of evidence that support existing beliefs or hypotheses (Nickerson 1998). Theory suggests that it is a category of implicit and unconscious processes that involve assembling one-sided evidence, and shaping it to fit one's belief. Equally important is that theory suggests these behaviors may be motivated or unmotivated, as one may selectively seek evidence in favor of a hypothesis, which one may or may not have a personal interest in supporting.

Hypotheses in present-day AI research are often implicit. Generally, these hypotheses are framed simply as a system being more accurate or efficient, compared to other systems. Where other fields, such as medicine or quantitative social sciences, would further articulate expectations in e.g. assigning specific conditions and considering effect sizes assigned to each competing hypothesis, in Computer Science and AI this is typically not done. This also may have to do with much of the published work being more of an engineering achievement, rather than a true hypothesis test seeking to explain and understand the world. However, in discussions on emerging qualities like AGI, this engineering positioning gets muddier, and more formal hypothesis testing would be justifiable: either one interprets outputs as in support of hints towards AGI (the alternative hypothesis), or as merely the result of an algorithm integrating qualities from the data it was trained on (the null hypothesis).

[5] https://ourworldindata.org/grapher/annual-scholarly-publications-on-artificial-intelligence?time=2010..2021

[6] Retrieved 23/01/23 using the search string "TITLE-ABS-KEY ((machine AND learning) OR (artificial AND intelligence) OR ai) AND PUBYEAR > 2009 AND PUBYEAR < 2024 " from the SCOPUS database.

Confirmation bias in hypothesis testing may manifest as a number of behaviors (Nickerson (1998)). Scientists may pay little to no attention to competing hypotheses or explanations, e.g. only considering the likelihood that outputs of a system support one's claims, and not the likelihood that the same outputs might occur if one's hypothesis is false. Similarly, bias may show when failing to articulate a sufficiently strong null hypothesis leading to a 'weak' or 'non-risky' experiment, a problem articulated in response to a number of scientific crises (Claesen et al. 2022). In extreme cases, propositions may be made that cannot be falsified based on how they are formulated. If the threshold to accept a favored hypothesis is too low, observations consistent with the hypothesis are almost guaranteed, and in turn fail to severely test the claim in question. Thus, one is far more likely to show evidence in favor of their beliefs by posing weak null hypotheses.

Related to the formulation of hypotheses is the interpretation of evidence in favor of competing hypotheses, wherein people will interpret identical evidence differently based on their beliefs. As Nickerson (1998) reviews, individuals may place greater emphasis or milder criticism on evidence in support of their hypothesis, and lesser emphasis and greater criticism on evidence that opposes it.

6.5. Outlook

Reflecting on the previous two sections, we make the following concrete recommendations for future research:

1. (*Acknowledgement of Human Bias*) Researchers should be mindful of, and explicit about, risks of human bias and anthropomorphization in interpreting results, which both can be done as part of the results discussion, but also in a dedicated 'limitations' section.

2. (*Stronger Testing*) Researchers should refrain from drawing premature conclusions about AGI, unless these are based on strong hypothesis tests.

3. (*Epistemologically Robust Standards*) We call for more precise definitions of terms like 'intelligence' and 'AGI', and publicly accountable and collaborative iterations over how we will measure them, with explicit room for independent reviewing and external auditing by the broader community.

Moreover, we believe that structural and cultural changes are in order to reduce current incentives to chase Big Statement Outcomes in AI research and industry. Our broadest and perhaps most ambitious goal is for our research community to **move away from authorship and instead embrace contributorship**. This argument has been raised long before in other research communities (Smith 1997) and more recently within our own (Liem and Demetriou 2023). Specifically, Liem and Demetriou (2023) argue that societally impactful scientific insights should be treated as open-source software artifacts. The open-source community sets a positive example of how scientific artifacts should be published in many different ways. Not only does

it adequately reward small contributions but it also naturally considers negative results (bugs) as part of the scientific process. Similarly, code reviews are considered so integral to the process that they typically end up as accredited contributions to projects. Open review platforms like OpenReview are a step in the right direction, but still fall short of what we know is technically feasible. Finally, software testing is, of course, not only essential but often obligatory before contributions are accepted and merged. As we have pointed out repeatedly in this work, any claims about AGI demand proper strong hypothesis tests. It is important to remember that AGI remains the alternative hypothesis and that the burden of proof therefore lies with those making strong claims.

6.6. CONCLUSION

As discussed above, AI research and development outcomes can easily be over-interpreted, both from a data perspective and because of human biases and interests. Academic researchers are not free from such biases. Thus, we call for the community to create explicit room for organized skepticism.

For research that seeks to explain a phenomenon, clear hypothesis articulation and strong null hypothesis formulation are needed. If claims of human-like or superhuman intelligence are made, these should be subject to severe tests (Claesen et al. 2022) that go beyond the display of surprise. Apart from focusing on getting novel improvements upon state-of-the-art published, organizing red-teaming activities as a community may help in incentivizing and normalizing constructive adversarial questioning. As the quest for AGI is so deeply rooted in human-like recognition, adding our voice to emerging calls to be vigilant in communication (Shanahan 2024), we put in an explicit word of warning about the use of terminology. Many terms used in current AGI research (e.g. emergence, intelligence, learning, 'better than human' performance) have a common understanding in specialized research communities, but have bigger, anthropomorphic connotations in laypersons. In fictional media, depictions of highly intelligent AI have for long been going around. In a study of films featuring robots, defined as "...an artificial entity that can sense and act as a result of (real-world or fictional) technology...", in the 134 most highly rated science-fiction movies on IMDB, 74 out of the 108 AI-robots studied had a humanoid shape, and 68 out of those had sufficient intelligence to interact at an almost human-level (Saffari et al. 2021). The authors identify human-like communication and the ability to learn as essential abilities in the depiction of AI agents in movies. They further show a common plot: humans perceive the AI agents as inferior, despite their possession of self-awareness and the desire to survive, which fuels the central conflict of the film, wherein humanity is threatened by AI superior in both intellect and physical abilities. It is often noted that experts and fictional content creators interact, informing and inspiring each other (Saffari et al. 2021; Neri and Cozman 2020).

This image also permeates present-day non-fictional writings on AI, which often use

anthropomorphized language (e.g. "ever more powerful digital minds" in the 'Pause Giant AI Experiments' open letter (Future of Life Institute 2023a)). In the news, we witness examples of humans falling in love with their AI chatbots (Morrone 2023; Steinberg 2023). The same news outlets discuss the human-like responses of Microsoft's Bing search engine, which had at that point recently been adopting GPT-4[7]. The article (Cost 2023), states "As if Bing wasn't becoming human enough" and goes on to claim it told them it loves them. Here, AI experts and influencers also have considerable influence on how the narrative unfolds on social media: according to Neri and Cozman (2020), actual AI-related harms did not trigger viral amplification (e.g. the death of an individual dying while a Tesla car was in autopilot, or the financial bankruptcy of a firm using AI technology to execute stock trades). Rather, potential risks expressed by someone perceived as having expertise and authority were amplified, such as statements made by Stephen Hawking during an interview in 2014.

We as academic researchers carry great responsibility for how the narrative will unfold, and what claims are believed. We call upon our colleagues to be explicitly mindful of this. As attractive as it may be to beat the state-of-the-art with a grander claim, let us return to the Mertonian norms, and thus safeguard our academic legitimacy in a world that only will be eager to run with made claims.

[7] A large multimodal language model from OpenAI https://openai.com/research/gpt-4.

7

CONCLUSION

Machine learning and artificial intelligence have developed rapidly in recent decades. Despite their success, state-of-the-art machine learning models are complex and their decision logic is difficult to interpret by humans. This thesis contributes to a growing body of research and literature that aims to tackle these issues and ultimately make opaque models more trustworthy. We have presented several technological innovations, methodological advances, empirical analyses and critical evaluations of existing paradigms and practices. In this final chapter, we conclude.

In Section 7.1, we begin by revisiting the main research questions set out at the beginning of this thesis. Next, we assess the real-world implications of this thesis and present an outlook for the future (Section 7.2). This is followed by a critical reflection on the limitations of our own work and potential threats to the validity of this thesis in Section 7.3. Finally, we present several recommendations for researchers and practitioners in Section 7.4.

7.1. REVISITING RESARCH QUESTIONS

In TRQ 1.1, we ask what counterfactual explanations are, why they might be useful for trustworthy AI and if there exist sufficient open-source implementations. Highlighting shortcomings of popular XAI approaches like LIME and SHAP, we argue that CE offer a useful and intuitive alternative. In particular, we explain that contrary to methods relying on local surrogate models, CE have full fidelity by construction as long as they are valid, which can be guaranteed (Guidotti 2022). Additionally, while CE can be manipulated much like LIME and SHAP, remedies are simple and cheap (Slack et al. 2020, 2021). In terms of other research contributions, our work in Chapter 2 also highlights a weakness of surrogate-based CE methods like *REVISE* (Joshi et al. 2019): the quality of the generated CE no longer depends exclusively on the quality of the opaque model, but also the surrogate. To the best of our knowledge, this is highlighted here for the first time. This observation has had considerable impact on our work in Chapter 4 and Chapter 5.

With respect to software availability, our study finds that—at the time it was first written—the availability of open-source software to explain opaque AI models through counterfactuals is still limited. While researchers have made piece-wise implementations of specific methods available[1], the only attempt at providing a unifying framework is the Python library CARLA (Pawelczyk et al. 2021). The Python library was released before CounterfactualExplanations.jl. It offers support for a number of different CE methods, but seems to no longer be actively maintained[2]. Within the Julia ecosystem, CounterfactualExplanations.jl is the first and only unifying framework for CE. Our work addresses a growing demand for packages that contribute towards trustworthy AI, which has been recognized and embraced by the Julia community. To the best of our knowledge, our framework is the only one that allows users to easily combine different CE methods and generate multiple counterfactuals in parallel using both multithreading and multiprocessing. This has enabled us to run experiments of a scale that is unprecedented in the field using one TU Delft's high-performance computing cluster ((DHPC) 2022).

Overall conclusion: Counterfactual explanations are an effective tool for trustworthy AI and our CounterfactualExplanations.jl package fills an important gap in the open-source software landscape.

In TRQ 1.2, we wonder about the dynamics of counterfactual explanation and algorithmic recourse, when they are implemented in practice. Our study finds that off-the-shelf counterfactual generators induce endogenous undesirable macrodynamics with respect to the underlying model and data, if not handled carefully. In particular, we show that a narrow focus on minimizing individual costs neglects the downstream effects of recourse itself, which carries real-world risks; a bank, for example, that offers individual recourse to its loan applicants, can be expected to face increased credit risk if minimal cost recourse is implemented. We find that independent of the application, classifier performance can be expected to deteriorate if models are retrained on datasets that include minimal cost counterfactuals.

A key observation is that minimal cost counterfactuals are fundamentally at odds with plausibility. As we explained already in the introduction, a counterfactual cannot be close to its factual starting point and close to the target domain both at the same time. Consequently, CE methods that target plausibility such as Joshi et al. (2019) and Schut et al. (2021), tend to suffer less from inducing undesirable dynamics than the baseline method by Wachter, Mittelstadt, and Russell (2017). We formalize this trade-off between individual costs and external costs due to implausibility and propose simple and effective mitigation strategies. An important consideration in this context is convergence: if the counterfactual search is discontinued immediately after the decision boundary is crossed, then it is unlikely the final counterfactual is plausible. In fact, we find that the simplest mitigation strategy for undesirable endogenous dynamics is to simply choose a high enough decision

[1] In addition to the examples listed in Chapter 15 of Molnar (2022) we have identified Schut et al. (2021)—a fast method for probabilistic classifiers—and Prado-Romero and Stilo (2022) for graph counterfactual explanations.

[2] At the time of writing, there has not been an update to the code base in over two years.

threshold for convergence: counterfactuals will typically end up deeper inside the target domain if the counterfactual search is considered as converged if and only if the classifier predicts the target class with high probability.

Overall conclusion: In the broader context of this thesis, the most important conclusion of Chapter 3 is that implausible counterfactuals can cause unexpected negative consequences in practice.

TRQ 1.3, asks if plausible explanations can be attained without relying on surrogate models. We demonstrate that it is indeed possible to rely exclusively on properties provided by the opaque model itself to achieve plausibility, but only if the model itself has actually learned plausible explanations for the data, where this latter condition is a feature of our proposed CE method, not a bug. To demonstrate this, our study begins by revisiting an observation from Chapter 2: using surrogate models to generate counterfactuals can affect the quality of the counterfactuals in unexpected and adverse ways. We provide a simple and yet compelling motivating example demonstrating that the surrogate-based *REVISE* generator (Joshi et al. 2019) can yield highly plausible counterfactuals, even if the opaque model has learned demonstrably implausible explanations for the data. Thus, we argue, it is possible to inadvertently "whitewash" an untrustworthy "black-box" model by effectively reallocating the task of learning plausible explanations from the model itself to the surrogate. To avoid such scenarios, we argue that inducing plausibility at all costs is a misguided paradigm. Instead, we should aim to generate counterfactuals that faithfully represent the conditional posterior distribution over inputs learned by the model.

To achieve this goal, we propose a new method for generating energy-constrained conformal counterfactuals—*ECCCo*. Our approach leverages ideas underlying joint energy-based models (Grathwohl et al. 2020) and conformal prediction. Specifically, we ensure that counterfactuals reach low-energy states with respect to the model and lead to high-certainty predictions of the target class. Through extensive experiments, we demonstrate the *ECCCo* achieves state-of-the-art levels of plausibility for well-specified models. This allows researchers and practitioners to use *ECCCo* to assess how trustworthy opaque models are based on the plausibility of the explanations they have learned.

Overall conclusion: Counterfactual explanations can be both plausible and faithful to the opaque model. Instead of aiming to develop model-agnostic tools for generating plausible explanations ("modelling explanations"), we should hold models accountable for delivering such explanations.

TRQ 1.4, asks a natural follow-up question: provided we have a way to generate faithful counterfactuals, can we use them to improve the trustworthiness of models? To this end, our work in Chapter 5 introduces a new training regime for differentiable models like artificial neural networks: counterfactual training. It involves generating faithful counterfactual explanations during each training iteration and then backpropagating model gradients with respect to the contrastive divergence

between counterfactuals and observed training samples in the target domain. Additionally, we interpret intermediate counterfactuals near the decision boundary as adversarial examples and penalize the model's adversarial loss. Our work therefore explicitly connects explainable to adversarial ML.

Our empirical findings demonstrate that CT yields more adversarially robust models that learn more plausible explanations for the data. Beyond plausibility and adversarial robustness, counterfactual training can also be used to ensure that models learn actionable explanations. To this end, we prove that CT induces models that are less sensitive to immutable or protected features. Importantly, our empirical results also show that these benefits with respect to trustworthiness do not come at the cost of reduced predictive performance. We find that predictive performance of models on test data is either unaffected by CT or more robust or both. The work in this chapter therefore provide substantial advances with respect to training more trustworthy AI.

Overall conclusion: Faithful counterfactual explanations can be leveraged during training to improve models with respect to explainability and adversarial robustness.

TRQ 1.5, enquires about the role of trustworthy AI in the context of LLMs. In Chapter 6, we critically assess a viral recent work that uses standard tools from mechanistic interpretability to arrive at the conclusion that modern LLMs learn world models. This in turn has been characterized as a milestone on the path towards AGI. Our study presents a number of experiments involving models of varying complexity to demonstrate that the finding of concept-related representations in latent spaces of models should not surprise us and certainly not be seen as evidence in favor of AGI. A thorough review of the social sciences' literature demonstrates why researchers might still fall into that trap, especially in an environment that has made AGI the north star of AI research (Blili-Hamelin et al. 2025). In summary, we caution researchers against misinterpreting results from mechanistic interpretability or else its role in the pursuit of more trustworthy AI may be tarnished.

Overall conclusion: Tools from mechanistic interpretability should be used carefully to avoid tarnishing their credibility with respect to trustworthy AI. Further work is needed to improve the usefulness of CE for LLMs.

7.2. IMPLICATIONS AND OUTLOOK

The findings in this work have shed light on many challenges and questions in the field of trustworthy AI and, in particular, counterfactual explanations. While we have also proposed solutions to some of the more specific challenges that we have encountered, our work highlights broader challenges that remain unanswered.

Chapter 3 and Chapter 4 have demonstrated that the field needs to rethink some of its core objectives. Chapter 3, in particular, showed that algorithmic recourse in practice involves multiple stakeholders typically competing for scarce resources.

As more opaque AI is deployed in the real world, we may have to rethink recourse as an economic and societal problem. Thus research on AR will inevitably inform future economic and societal questions and vice versa. Our findings from Chapter 4 require us to rethink what we truly want to get out of XAI methods: practical but possibly misleading answers or enhanced understanding of model behavior? Our findings also invite follow-up questions about the evaluation of counterfactual explanations. While we provide a nuanced definition and metric for faithfulness, we do not pretend to provide final answers in Chapter 4 and believe that this objective for counterfactuals deserves more attention.

Chapter 5 has important implications for the connection between explainability and adversarial robustness in machine learning. Our framework for counterfactual training constitutes a solid starting ground, but there is likely much untapped potential for synergies between these two subfields of trustworthy AI. We therefore believe that researchers from both communities would benefit from collaborating. We further believe that practitioners would benefit from taking a holistic approach to trustworthy AI, explicitly recognizing that various objectives may complement each other but also compete.

In all of this, we hope that the work presented in Chapter 2 can continue to play a role in facilitating research and experimentation. The broader ecosystem of packages that have grown out of this initial work have certainly gained some traction and popularity in the Julia community, but to create a lasting impact they will need to continue to be maintained and developed further. We believe that Taija has great potential to for both research and industry.

Finally, results presented in this thesis also have implications for the ongoing discourse around AGI. Chapter 6 has shown that we should insist on adhering to scientific principles when engaging in this discourse as academics, especially those among us who are considered as thought leaders by many. As a whole, this thesis has also shown that we are still struggling to truly understand and control the behavior of even the most basic building blocks of AI. This should give anyone in our field currently treating AGI as the north-star goal of AI research serious pause.

To end on an optimistic note, we believe that this work also provides hope for trustworthy AI. We have shown that it is possible to use model explanations for good: if carefully constructed, they can help us to not only assess the trustworthiness of opaque models, but also improve it. This requires work, but as economists like to say: "There is no such thing as a free lunch"[3].

[3]This quote is often attributed to Milton Friedman, but it likely originated earlier. According to the Cambridge Dictionary, the phrase is used to emphasize that you cannot get something for nothing: https://dictionary.cambridge.org/us/dictionary/english/there-s-no-such-thing-as-a-free-lunch

7.3. LIMITATIONS AND THREATS TO VALIDITY

In this section, we highlight limitations and threats to the validity of our work. We focus on points that were not already discussed explicitly in the context of individual chapters.

7.3.1. CONSTRUCT VALIDITY

Our evaluations of counterfactuals in Chapter 4 and Chapter 5 rely on imperfect metrics used to assess the plausibility and faithfulness of explanations. For plausibility, we extend existing distance-based metrics for measuring the dissimilarity between counterfactuals and observed training data in the target domain. This is a valid approach to assess plausibility, to the extent that "broadly consistent with the observed data" is an adequate proxy for "plausible as assessed by humans". We found in our own work that this is not always the case: in Chapter 4, for instance, we found that image counterfactuals produced by *ECCCo* were sometimes more visually appealing and plausible than the distance-based evaluation metrics suggested. To mitigate this, we tested different distance measures and eventually introduced a new divergence measure in Chapter 5, but we recognize that ideally plausibility of explanations would be assessed directly by humans. The scale of our experiments involving multiple millions of counterfactuals, made this option infeasible.

Regarding faithfulness, we rely on established methods for estimating the conditional model posterior over inputs (Grathwohl et al. 2020; Murphy 2023). This approach has two potential shortcomings with respect to the validity of our work: firstly, the estimated empirical distributions are subject to estimation error; secondly, our proposed method in Chapter 4 is biased towards this metric by construction, although the same can be said about other methods targeting certain metrics like minimal costs. The important finding in this context is that our proposed counterfactual generator can satisfy its primary target (faithfulness) while also achieving its secondary target (plausibility).

7.3.2. INTERNAL VALIDITY

Further on evaluation, we rely on cross-validation to account for stochasticity: specifically, we always generate and evaluate counterfactuals multiple times, each time drawing a different random subsample of the available data. We then compute averages and standard deviations of our evaluation metrics to get a sense of how substantial and significant the differences in outcomes are for the various methods we test. This is consistent with common practice in the related literature and—we believe—sufficient to arrive at the conclusions we present in individual chapters. Nonetheless, we recognize that in Chapter 4 we fall short of testing our evaluations and rankings as rigorously for statistical significance as we do in Chapter 3.

The internal validity of the findings in Chapter 3 is largely threatened by the simplifying assumptions we make about stakeholder actions. For example, we assume that any individual provided with a valid algorithmic recourse ends up implementing the exact recommendations. We also assume that model owners retrain models regularly after individuals have implemented recourse and that no entirely new samples are added to the training population. Retraining is continued over multiple rounds even in the face of model deterioration and other negative dynamics after the first few rounds. All of these modelling assumptions are necessarily simple to focus on our main narrative. We intentionally abstract from detail to study the worst-case high-level effects we are interested in.

7.3.3. EXTERNAL VALIDITY

To empirically test our claims and proposed methods in Chapter 3, Chapter 4 and Chapter 5, we employed both synthetic and publicly available real-world datasets that are commonly studied in the related literature. We have also largely relied on studying small and simple neural network architectures, again consistent with the related literature. While we have made an effort to always include a broad range of sources to ensure a certain degree of robustness in our findings, it is certainly possible and indeed expected that some of our findings do not always hold true in practice. We expect this in some cases, because certain results are subject to hyperparameter sensitivity, in particular results from Chapter 4 and to a lesser degree Chapter 5. A related threat to external validity is scalability: the computational cost involved in generating counterfactual explanations increases in the dimensionality of inputs, which may make certain methods we propose—in particular counterfactual training—computationally prohibitive.

Concerning Chapter 6, it could be argued that the experiments we present involve models that are too simple to warrant any discussion around AGI[4]. Our response to this would be that the choice of simple models that have not previously been linked to AGI is very much intentional. As our experiments show, properties of LLMs that have been presented as novel and surprising are in fact shared by much simpler models.

7.3.4. SOFTWARE LIMITATIONS

Since Chapter 2 was published in 2023, we have continued to actively develop and maintain CounterfactualExplanations.jl, such that it has now reached maturity with respect to fundamental features. That being said, to the best of our knowledge it has never been tested in a production environment involving larger models and datasets than the ones we have used in our research. Due to our focus on simulations and cross-validations involving many counterfactuals (high

[4]This is in fact how one of the few dissenting audience members at ICML dismissed our work without any further consideration.

n) of typically moderate dimensionality (small p), we have prioritized support for parallelization through multithreading and multiprocessing, as opposed to graphical processing units (GPUs). Thankfully, Julia offers fantastic support also for the latter and since we rely on standard routines for autodifferention, it should be straightforward to address this limitation. Beyond this, there are numerous smaller outstanding development tasks listed on our repository: JuliaTrustworthyAI/CounterfactualExplanations.jl/issues.

Concerning internal validity, our software is no different from any other software in that it contains bugs and inefficiencies. We have encountered such shortcomings in the past and expect to find more in the future. Relatedly, certain software architecture and prioritization choices we have made may be suboptimal for specific applications, even though they have served us well in the past. Regarding external validity, there is a strong possibility that Patrick will not be able to maintain it as actively in the past. To address this risk, we have taken steps to attract external contributions in the past and aim to continue in this fashion.

7.4. RECOMMENDATIONS FOR RESEARCH AND PRACTICE

In this section, we provide general recommendations for both researchers and practitioners working with opaque machine learning models. They are derived from our research findings but not in all cases directly tied to specific results.

Recommendation 7.1

Beware of high-dimensional spaces, especially large latent parameter spaces of machine learning models.

As proposition (5) of this thesis states: strange things really do happen in high-dimensional spaces. This is an observation that universally applies to all chapters of this thesis in one way or another. Recommendation 7.1 can be seen as general call for caution. Even though we—along with many others working in the field—have contributed through our work towards making opaque models *more* trustworthy, there is simply no silver bullet. For better or worse, high degrees of freedom in representation learning make models susceptible to learning representations that humans cannot interpret. This is what makes such models so powerful at achieving narrow objectives. But as we have seen throughout this thesis, it also has the potential to make them sensitive to spurious associations in the data they are trained on. Our work contributes several results that can aid researchers in navigating this challenge, but we want to be very clear that we think of our findings as remedies, not a cure.

We consider explainable and trustworthy AI as a moving target, just like adversarial robustness is still considered an unsolved challenge even for simple models (Kolter 2023): for every explanation and any attack we have identified, another is likely to

follow. Thus, we recommend that researchers and practitioners avoid striving for trustworthy AI as some attainable end goal, and instead recognize that its continuous process that requires work. Counterfactual explanations provide a particularly useful framework to deal with models that enjoy large degrees of freedom, precisely because they are similarly unconstrained in terms of the feature space they can occupy.

Recommendation 7.2

Explanations are rarely unique and researchers and practitioners studying and using opaque machine learning models should embrace this fact.

Multiplicity of counterfactual explanations is a feature, not a bug. Uniqueness has in the past been considered as an explicit goal in the context of XAI, possibly because humans are naturally inclined to prefer straight answers over complicated ones. We would argue, however, that the notion of finding unique solutions to our search for model explanations is fundamentally at odds with basic properties of the models we are studying: they are not unique solutions either, not even to the narrow objectives we typically train them for. Any fitted neural network is just a random outcome of a stochastic training process that could have resulted in any one of many different paramterizations that provide compelling explanations for the data (Wilson 2020).

More to the point of explainability and its use cases, our work has also frequently shown that real-world objectives are not in fact narrow: Chapter 3 highlighted the trade-off between individual and external costs in algorithmic recourse; Chapter 4 shed light on the interplay between plausibility and faithfulness of counterfactual explanations; and in Chapter 5 we have made the case for explicitly adjusting training objectives to induce models to learn actionable and plausible explanations. Since objectives are multiple and context-dependent, explanations are also inevitably variable. Therefore, it is our recommendation that researchers and practitioners use tools for trustworthy AI that are flexible enough to accommodate such multiplicity of objectives and explanations.

Recommendation 7.3

Explanations for models should be faithful first, plausible second.

Speaking of objectives, we believe that the guiding objective for counterfactual explanations—and in fact any XAI method—should be faithfulness to the model. It is very difficult to think of scenarios that call for plausible, robust, diverse, actionable or easily affordable recourse recommendations that do not also faithfully explain the model in question. At best, we would consider this a short-term solution to dealing with opaque models in practice. As we have demonstrated in Chapter 4, it is entirely possible to generate plausible explanations for accurate and yet fundamentally untrustworthy models. This should be keenly avoided, since it may instill trust in models that are not worthy of it.

Recommendation 7.4

Hold models accountable for learning plausible explanations, instead of modelling explanations to your liking.

In close relation to Recommendation 7.3, we recommend that researchers in XAI avoid considering plausible, model-agnostic explanations as the holy grail. Explanations do not make automated decisions that may affect the lives of individuals, models do. Explanations are merely reflections of how models arrive at the decisions they make. Thus, we should use explanations primarily to inform our understanding of models and strive to improve models based on explanations, instead of treating models as fixed and tailoring our explanation around them. Otherwise, we risk still treating models as oracles that cannot be held accountable, much like we have done in the past (O'Neil 2016).

This is more important in the age of LLMs than ever before. As we have argued in Chapter 6, people are prone to anthropomorphize and idolize complex technologies they do not fully understand. There is a real risk today that people are so dumbstruck and overwhelmed by machines that are quite literally optimized to appeal to them, that people end up blindly relying on them, even worshiping them. This, coupled with a lack of accountability, provides model owners with unprecedented powers to affect individuals. No matter how powerful these models become, we need to avoid thinking that they are inscrutable, leave alone infallible.

Recommendation 7.5

Invite diverse perspectives into research and practice.

This last Recommendation 7.5 is not directly tied to any specific result of this work, but rather the thesis as a whole and the direct or indirect contributions by the many people that co-shaped it. In times when diversity is once again under threat by narrow-minded people in powerful positions, we find it important to share that in our own experience, diverse perspectives supercharge innovation. Many of the findings in this thesis have resulted from combining ideas from different subfields of AI or even external domains including economics and other social sciences. This culminated in Chapter 6, which involved co-authors from a variety of different disciplines and is likely going to be one of the more influential contributions of this thesis.

REFERENCES

Abbasnejad, Ehsan, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. "Counterfactual Vision and Language Learning." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10041–51. https://doi.org/10.1109/CVPR42600.2020.01006.

Ackerman, Samuel, Parijat Dube, Eitan Farchi, Orna Raz, and Marcel Zalmanovici. 2021. "Machine Learning Model Drift Detection Via Weak Data Slices." In *2021 IEEE/ACM Third International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)*, 1–8. IEEE. https://doi.org/10.1109/deeptest52559.2021.00007.

Agustí, Marc, Ignacio Vidal-Quadras Costa, and Patrick Altmeyer. 2023. "Deep Vector Autoregression for Macroeconomic Data." *IFC Bulletins Chapters* 59. https://www.bis.org/ifc/publ/ifcb59_39.pdf.

Alain, Guillaume, and Yoshua Bengio. 2016. "Understanding intermediate layers using linear classifier probes." *ArXiv*. https://api.semanticscholar.org/CorpusID:9794990.

Alperin, Juan P, Carol Muñoz Nieves, Lesley A Schimanski, Gustavo E Fischman, Meredith T Niles, and Erin C McKiernan. 2019. "How significant are the public dimensions of faculty work in review, promotion and tenure documents?" *ELife* 8: e42254.

Altman, Sam. 2025. "Reflections." January 6, 2025. https://blog.samaltman.com/reflections.

Altmeyer, Patrick, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Endogenous Macrodynamics in Algorithmic Recourse." IEEE. https://doi.org/10.1109/satml54575.2023.00036.

Altmeyer, Patrick, Leva Boneva, Rafael Kinston, Shreyosi Saha, and Evarist Stoja. 2023. "Yield Curve Sensitivity to Investor Positioning Around Economic Shocks." Bank of England working papers 1029. Bank of England. https://doi.org/None.

Altmeyer, Patrick, Aleksander Buszydlik, Arie van Deursen, and Cynthia C. S. Liem. 2026. "Counterfactual Training: Teaching Models Plausible and Actionable Explanations." https://arxiv.org/abs/2601.16205.

Altmeyer, Patrick, Andrew M Demetriou, Antony Bartlett, and Cynthia C. S. Liem. 2024. "Position: Stop Making Unscientific AGI Performance Claims." In *International Conference on Machine Learning*, 1222–42. PMLR. https://proceedings.mlr.press/v235/altmeyer24a.html.

Altmeyer, Patrick, Arie van Deursen, and Cynthia C. S. Liem. 2023b. "Explaining Black-Box Models through Counterfactuals." In *Proceedings of the JuliaCon Conferences*, 1:130. https://doi.org/10.21105/jcon.00130.

———. 2023a. "Explaining Black-Box Models through Counterfactuals." In *Proceedings of the JuliaCon Conferences*, 1:130. https://doi.org/10.21105/jcon.00130.

Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2024a. "Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals." In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 38:10829–37. 10. https://doi.org/10.1609/aaai.v38i10.28956.

———. 2024b. "Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals." In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 38:10829–37. 10. https://doi.org/10.1609/aaai.v38i10.28956.

Altmeyer, Patrick, Pedro Gurrola-Perez, Rafael Kinston, and Jessica Redmond. 2019. "Modelling the Demand for Central Bank Reserves." https://www.ecb.europa.eu/press/conferences/html/20191111_ecb_money_market_workshop_conference.en.html.

Angelopoulos, Anastasios N., and Stephen Bates. 2022. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." https://arxiv.org/abs/2107.07511.

Antorán, Javier, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2020. "Getting a Clue: A Method for Explaining Uncertainty Estimates." https://arxiv.org/abs/2006.06848.

Arcones, Miguel A, and Evarist Gine. 1992. "On the Bootstrap of U and V Statistics." *The Annals of Statistics*, 655–74.

Arrieta, Alejandro Barredo, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Artelt, André, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. "Evaluating Robustness of Counterfactual Explanations." In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–09. IEEE.

Augustin, Maximilian, Alexander Meinke, and Matthias Hein. 2020. "Adversarial Robustness on In- and Out-Distribution Improves Explainability." In *Computer Vision – ECCV 2020*, edited by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, 228–45. Cham: Springer.

Balashankar, Ananth, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. 2023. "Improving Classifier Robustness through Active Generative Counterfactual Data Augmentation." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 127–39. ACL. https://doi.org/10.18653/v1/2023.findings-emnlp.10.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2022. "Fairness and Machine Learning." December 2022. https://fairmlbook.org/index.html.

Barry Becker, Ronny Kohavi. 1996. "Adult." UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Becker, Barry, and Ronny Kohavi. 1996. "Adult." UCI Machine Learning Repository.

Belinkov, Yonatan. 2021. "Probing Classifiers: Promises, Shortcomings, and Advances." https://arxiv.org/abs/2102.12452.

Bell, Andrew, Joao Fonseca, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. 2024. "Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity." https://arxiv.org/abs/2401.16088.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Berardi, Andrea, and Alberto Plazzi. 2022. "Dissecting the yield curve: The international evidence." *Journal of Banking & Finance* 134: 106286.

Bereska, Leonard, and Efstratios Gavves. 2024. "Mechanistic Interpretability for AI Safety – a Review." https://arxiv.org/abs/2404.14082.

Berlinet, Alain, and Christine Thomas-Agnan. 2011. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-9096-9.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98. https://doi.org/10.1137/141000671.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. "The Values Encoded in Machine Learning Research." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*.

Bischl, Bernd, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, et al. 2023. "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges." *WIREs Data Mining and Knowledge Discovery* 13 (2): e1484. https://doi.org/https://doi.org/10.1002/widm.1484.

Blaom, Anthony D., Franz Kiraly, Thibaut Lienart, Yiannis Simillides, Diego Arenas, and Sebastian J. Vollmer. 2020. "MLJ: A Julia Package for Composable Machine Learning." *Journal of Open Source Software* 5 (55): 2704. https://doi.org/10.21105/joss.02704.

Blili-Hamelin, Borhane, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, et al. 2025. "Position: Stop treating 'AGI' as the north-star goal of AI research." In *Forty-Second International Conference on Machine Learning Position Paper Track*.

Borch, Christian. 2022. "Machine Learning, Knowledge Risk, and Principal-Agent Problems in Automated Trading." *Technology in Society*, 101852. https://doi.org/10.1016/j.techsoc.2021.101852.

Borisov, Vadim, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. "Deep Neural Networks and Tabular Data: A Survey." *IEEE Transactions on Neural Networks and Learning Systems*.

Brunnermeier, Markus K. 2016. "Bubbles." In *Banking Crises: Perspectives from*

the New Palgrave Dictionary, 28–36. Springer.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Buszydlik, Aleksander, Patrick Altmeyer, Cynthia C. S. Liem, and Roel Dobbe. 2024. "Grounding and Validation of Algorithmic Recourse in Real-World Contexts: A Systematized Literature Review." https://openreview.net/pdf?id=oEmyoy5H5P.

———. 2025. "Understanding the Affordances and Constraints of Explainable AI in Safety-Critical Contexts: A Case Study in Dutch Social Welfare." In *Electronic Government. EGOV 2025. Lecture Notes in Computer Science*. upcoming.

Carlisle, M. 2019. "Racist Data Destruction? - a Boston Housing Dataset Controversy." https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8.

Carrizosa, Emilio, Jasone Ramırez-Ayerbe, and Dolores Romero. 2021. "Generating Collective Counterfactual Explanations in Score-Based Classification via Mathematical Optimization."

Caterino, Pasquale. 2024. "Google Summer of Code 2024 Final Report: Add Support for Conformalized Bayes to ConformalPrediction.jl." https://gist.github.com/pasq-cat/f25eebc492366fb6a4f428426f93f45f.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. "Anomaly Detection: A Survey." *ACM Computing Surveys (CSUR)* 41 (3): 1–58.

Claesen, Aline, Daniel Lakens, Noah van Dongen, et al. 2022. "Severity and Crises in Science: Are We Getting It Right When We're Right and Wrong When We're Wrong?"

Cost, Ben. 2023. "Bing AI chatbot goes on 'destructive' rampage: 'I want to be powerful — and alive'." https://nypost.com/2023/02/16/bing-ai-chatbots-destructive-rampage-i-want-to-be-powerful/.

Crump, Richard K, and Nikolay Gospodinov. n.d. "Deconstructing the yield curve."

Dandl, Susanne, Andreas Hofheinz, Martin Binder, Bernd Bischl, and Giuseppe Casalicchio. 2023. "Counterfactuals: An R Package for Counterfactual Explanation Methods." arXiv. http://arxiv.org/abs/2304.06569.

Dasgupta, Sanjoy. 2013. "Experiments with Random Projection." https://arxiv.org/abs/1301.3849.

Delaney, Eoin, Derek Greene, and Mark T. Keane. 2021. "Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions." arXiv. http://arxiv.org/abs/2107.09734.

(DHPC), Delft High Performance Computing Centre. 2022. "DelftBlue Supercomputer (Phase 1)." https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1.

Dhurandhar, Amit, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. "Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." *Advances in Neural Information Processing Systems* 31.

Dombrowski, Ann-Kathrin, Jan E Gerken, and Pan Kessel. 2021. "Diffeomorphic Explanations with Normalizing Flows." In *ICML Workshop on Invertible Neural*

Networks, Normalizing Flows, and Explicit Likelihood Models.

Du, Yilun, and Igor Mordatch. 2020. "Implicit Generation and Generalization in Energy-Based Models." https://arxiv.org/abs/1903.08689.

Epley, Nicholas, Adam Waytz, and John T Cacioppo. 2007. "On seeing human: a three-factor theory of anthropomorphism." *Psychological Review* 114 (4): 864.

Falk, Ruma, and Clifford Konold. 1997. "Making sense of randomness: Implicit encoding as a basis for judgment." *Psychological Review* 104 (2): 301.

Fan, Fenglei, Jinjun Xiong, and Ge Wang. 2020. "On Interpretability of Artificial Neural Networks." https://arxiv.org/abs/2001.02522.

Field, Hayden. 2025. "OpenAI Partners with u.s. National Laboratories on Scientific Research, Nuclear Weapons Security." CNBC. January 30, 2025. https://www.cnbc.com/2025/01/30/openai-partners-with-us-national-laboratories-on-scientific-research.html.

Foster, Kevin R, and Hanna Kokko. 2009. "The evolution of superstitious and superstition-like behaviour." *Proceedings of the Royal Society B: Biological Sciences* 276 (1654): 31–37.

Frankle, Jonathan, and Michael Carbin. 2019. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." In *International Conference on Learning Representations.*

Freiesleben, Timo. 2022. "The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples." *Minds and Machines* 32 (1): 77–109.

Future of Life Institute. 2023a. "Pause Giant AI Experiments: An Open Letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

———. 2023b. "Pause Giant AI Experiments: An Open Letter." March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. "A Survey on Concept Drift Adaptation." *ACM Computing Surveys (CSUR)* 46 (4): 1–37.

Goertzel, Ben. 2014. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5 (1): 1.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press.

Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. 2015. "Explaining and Harnessing Adversarial Examples." https://arxiv.org/abs/1412.6572.

Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. "Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One." In *International Conference on Learning Representations.*

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. "A Kernel Two-Sample Test." *The Journal of Machine Learning Research* 13 (1): 723–73.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" https://arxiv.org/abs/2207.08815.

Guidotti, Riccardo. 2022. "Counterfactual Explanations and How to Find Them:

Literature Review and Benchmarking." *Data Mining and Knowledge Discovery* 38 (5): 2770–2824. https://doi.org/10.1007/s10618-022-00831-6.

Guo, Hangzhi, Thanh H. Nguyen, and Amulya Yadav. 2023. "CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations." In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 577--589. KDD '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3580305.3599290.

Gurnee, Wes, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. "Finding Neurons in a Haystack: Case Studies with Sparse Probing." *arXiv Preprint arXiv:2305.01610*.

Gurnee, Wes, and Max Tegmark. 2023a. "Language Models Represent Space and Time." *arXiv Preprint arXiv:2310.02207v2*.

———. 2023b. "Language Models Represent Space and Time." *arXiv Preprint arXiv:2310.02207v1*.

Haig, Brian D. 2003. "What is a spurious correlation?" *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences* 2 (2): 125–32.

Haltmeier, Markus, and Linh Nguyen. 2023. "Regularization of Inverse Problems by Neural Networks." In *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, 1065–93. Springer.

Hanneke, Steve. 2007. "A Bound on the Label Complexity of Agnostic Active Learning." In *Proceedings of the 24th International Conference on Machine Learning*, 353–60. https://doi.org/10.1145/1273496.1273541.

Heider, Fritz, and Marianne Simmel. 1944. "An experimental study of apparent behavior." *The American Journal of Psychology* 57 (2): 243–59.

Hengst, Floris, Ralf Wolter, Patrick Altmeyer, and Arda Kaygan. 2024. "Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition." In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2412–32. https://doi.org/10.18653/v1/2024.findings-naacl.156.

Hoffman, Hans. 1994. "German Credit Data." https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Innes, Michael, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. 2018. "Fashionable Modelling with Flux." https://arxiv.org/abs/1811.01457.

Innes, Mike. 2018. "Flux: Elegant Machine Learning with Julia." *Journal of Open Source Software* 3 (25): 602. https://doi.org/10.21105/joss.00602.

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." https://arxiv.org/abs/1907.09615.

Kaggle. 2011. "Give Me Some Credit, Improve on the State of the Art in Credit Scoring by Predicting the Probability That Somebody Will Experience Financial Distress in the Next Two Years." https://www.kaggle.com/c/GiveMeSomeCredit; Kaggle. https://www.kaggle.com/c/GiveMeSomeCredit.

Karimi, Amir-Hossein, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2021.

"A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects." https://arxiv.org/abs/2010.04050.

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. 2021. "Algorithmic Recourse: From Counterfactual Explanations to Interventions." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–62. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445899.

Karimi, Amir-Hossein, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. "Algorithmic Recourse Under Imperfect Causal Knowledge: A Probabilistic Approach." https://arxiv.org/abs/2006.06831.

Kaufmann, Maximilian, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Nicolas Papernot. 2022. "Efficient Adversarial Training with Data Pruning." *arXiv Preprint arXiv:2207.00694*.

Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219.

Kingma, Diederik P., and Jimmy Ba. 2017. "Adam: A Method for Stochastic Optimization." https://arxiv.org/abs/1412.6980.

Kıcıman, Emre, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. "Causal Reasoning and Large Language Models: Opening a New Frontier for Causality." *arXiv Preprint arXiv:2305.00050*.

Kloft, Agnes Mercedes, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. ""AI Enhances Our Performance, i Have No Doubt This One Will Do the Same": The Placebo Effect Is Robust to Negative Descriptions of AI." In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24.

Kolter, Zico. 2023."Keynote Addresses: SaTML 2023 ." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/SaTML54575.2023.00009.

Krizhevsky, A. 2009. "Learning Multiple Layers of Features from Tiny Images." In. https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

Kumar, Sumit. 2022. "Effective hedging strategy for us treasury bond portfolio using principal component analysis." *Academy of Accounting and Financial Studies* 26 (1).

Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. 2017. "Adversarial Machine Learning at Scale." https://arxiv.org/abs/1611.01236.

Ladouceur, Robert, Claude Paquet, and Dominique Dubé. 1996. "Erroneous Perceptions in Generating Sequences of Random Events 1." *Journal of Applied Social Psychology* 26 (24): 2157–66.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6405–16. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.

Laugel, Thibault, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. "Inverse Classification for Comparison-Based Interpretability in Machine Learning." arXiv. https://doi.org/10.48550/arXiv.1712.08443.

LeCun, Yann. 1998. "The MNIST database of handwritten digits." http://yann.lecun.com/exdb/mnist/.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–2324.

Leofante, Francesco, and Nico Potyka. 2024. "Promoting Counterfactual Robustness Through Diversity." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (19): 21322–30. https://doi.org/10.1609/aaai.v38i19.30127.

Li, Kenneth, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." *arXiv Preprint arXiv:2210.13382*.

Liem, Cynthia C. S., and Andrew M Demetriou. 2023. "Treat Societally Impactful Scientific Insights as Open-Source Software Artifacts." In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 150–56. IEEE.

Lippe, Phillip. 2024. "UvA Deep Learning Tutorials." https://uvadlc-notebooks.readthedocs.io/en/latest/.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." https://arxiv.org/abs/1907.11692.

Luiz Franco, Jorge. 2024. "JSoC: When Causality Meets Recourse." https://www.taija.org/blog/posts/causal-recourse/.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77.

Luu, Hoai Linh, and Naoya Inoue. 2023. "Counterfactual Adversarial Training for Improving Robustness of Pre-trained Language Models." In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 881–88. ACL. https://aclanthology.org/2023.paclic-1.88/.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. "Towards Deep Learning Models Resistant to Adversarial Attacks." *arXiv Preprint arXiv:1706.06083*.

Mahajan, Divyat, Chenhao Tan, and Amit Sharma. 2020. "Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers." https://arxiv.org/abs/1912.03277.

Manokhin, Valery. 2022. "Awesome Conformal Prediction." Zenodo. https://doi.org/10.5281/zenodo.6467205.

Marcus, Gary. 2023. "Muddles about Models." https://garymarcus.substack.com/p/muddles-about-models.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina

Ligett, Terah Lyons, James Manyika, et al. 2023. "Artificial Intelligence Index Report 2023." Institute for Human-Centered AI.

McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. 2012. "Big Data: The Management Revolution." *Harvard Business Review* 90 (10): 60–68.

Merton, Robert K et al. 1942. "Science and technology in a democratic order." *Journal of Legal and Political Sociology* 1 (1): 115–26.

Michal S Gal, Daniel L Rubinfeld. 2019. "Data Standardization." *NYUL Rev.*

Miller, John, Smitha Milli, and Moritz Hardt. 2020. "Strategic Classification Is Causal Modeling in Disguise." In *Proceedings of the 37th International Conference on Machine Learning*, 6917–26. PMLR. https://proceedings.mlr.press/v119/miller20b.html.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Mishkin, Frederic S et al. 2008. "How Should We Respond to Asset Price Bubbles." *Financial Stability Review* 12 (1): 65–74.

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

Morrone, Megan. 2023. "Replika exec: AI friends can improve human relationships." https://www.axios.com/2023/11/09/replika-blush-rita-popova-ai-relationships-dating.

Mothilal, Ramaravind K, Amit Sharma, and Chenhao Tan. 2020. "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–17. https://doi.org/10.1145/3351095.3372850.

Müller, Petra, and Matthias Hartmann. 2023. "Linking paranormal and conspiracy beliefs to illusory pattern perception through signal detection theory." *Scientific Reports* 13 (1): 9739.

Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.

———. 2023. *Probabilistic Machine Learning: Advanced Topics*. MIT press.

Nanda, Neel, Andrew Lee, and Martin Wattenberg. 2023. "Emergent Linear Representations in World Models of Self-Supervised Sequence Models." *arXiv Preprint arXiv:2309.00941*.

Nelson, Kevin, George Corbin, Mark Anania, Matthew Kovacs, Jeremy Tobias, and Misty Blowers. 2015. "Evaluating Model Drift in Machine Learning Algorithms." In *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1–8. IEEE. https://doi.org/10.1109/cisda.2015.7208643.

Neri, Hugo, and Fabio Cozman. 2020. "The role of experts in the public perception of risk of artificial intelligence." *AI & SOCIETY* 35: 663–73.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology* 2 (2): 175–220.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Oliveira, Raphael Mazzine Barbosa de, and David Martens. 2021. "A Framework and Benchmarking Study for Counterfactual Generating Methods on Tabular Data." *Applied Sciences* 11 (16): 7274. https://doi.org/10.3390/app11167274.

OM, ANDERS BJ ORKSTR. 2001. "Ridge Regression and Inverse Problems." *Stockholm University, Department of Mathematics*.

OpenAI. 2025. "Strengthening America's AI Leadership with the u.s. National Laboratories." Edited by OpenAI. January 30, 2025. https://openai.com/index/strengthening-americas-ai-leadership-with-the-us-national-laboratories/.

Pace, R Kelley, and Ronald Barry. 1997. "Sparse Spatial Autoregressions." *Statistics & Probability Letters* 33 (3): 291–97. https://doi.org/10.1016/s0167-7152(96)00140-x.

Pawelczyk, Martin, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. "Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis." In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, edited by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, 151:4574–94. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v151/pawelczyk22a.html.

Pawelczyk, Martin, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 2021. "CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms." https://arxiv.org/abs/2108.00783.

Pawelczyk, Martin, Teresa Datta, Johannes van-den-Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. 2023. "Probabilistically Robust Recourse: Navigating the Trade-Offs Between Costs and Robustness in Algorithmic Recourse." https://arxiv.org/abs/2203.06768.

Pindyck, Robert S, and Daniel L Rubinfeld. 2014. *Microeconomics*. Pearson Education.

Poyiadzi, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. "FACE: Feasible and Actionable Counterfactual Explanations." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–50.

Prado-Romero, Mario Alfonso, and Giovanni Stilo. 2022. "GRETEL: Graph Counterfactual Explanation Evaluation Framework." In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. CIKM '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3511808.3557608.

Prince, Simon J. D. 2023. *Understanding Deep Learning*. The MIT Press. http://udlbook.com.

Rabanser, Stephan, Stephan Günnemann, and Zachary Lipton. 2019. "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift." *Advances in Neural Information Processing Systems* 32.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 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, 1135–44.

Rooij, Iris van, Olivia Guest, Federico G Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. 2023. "Reclaiming AI as a theoretical tool for cognitive science." *psyarXiv*. https://osf.io/4cbuv.

Ross, Alexis, Himabindu Lakkaraju, and Osbert Bastani. 2024. "Learning Models for Actionable Recourse." In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Saffari, Ehsan, Seyed Ramezan Hosseini, Alireza Taheri, and Ali Meghdari. 2021. "'Does cinema form the future of robotics?': a survey on fictional robots in sci-fi movies." *SN Applied Sciences* 3 (6): 655.

Sauer, Axel, and Andreas Geiger. 2021. "Counterfactual Generative Networks." https://arxiv.org/abs/2101.06046.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2024. "Are Emergent Abilities of Large Language Models a Mirage?" *Advances in Neural Information Processing Systems* 36.

Schut, Lisa, Oscar Key, Rory McGrath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. "Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties." In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.

Shah, Agam, Suvan Paturi, and Sudheer Chava. 2023. "Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis." *arXiv Preprint arXiv:2310.02207v1*. https://arxiv.org/abs/2305.07972.

Shanahan, Murray. 2024. "Talking about Large Language Models." *Communications of the ACM* 67 (2): 68–79.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. 2020. "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 166–72. AIES '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375812.

Slack, Dylan, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. "Counterfactual Explanations Can Be Manipulated." *Advances in Neural Information Processing Systems* 34.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. "Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–86.

Smith, Richard. 1997. "Authorship Is Dying: Long Live Contributorship: The BMJ Will Publish Lists of Contributors and Guarantors to Original Articles." *Bmj*. British Medical Journal Publishing Group.

Spooner, Thomas, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and

Daniele Magazzeni. 2021. "Counterfactual Explanations for Arbitrary Regression Models." https://arxiv.org/abs/2106.15212.

Steinberg, Brooke. 2023. "I fell in love with an AI chatbot — she rejected me sexually." https://nypost.com/2023/04/03/40-year-old-man-falls-in-love-with-ai-chatbot-phaedra/.

Stutz, David, Krishnamurthy, Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. 2022. "Learning Optimal Conformal Classifiers." https://arxiv.org/abs/2110.09192.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." https://arxiv.org/abs/1703.01365.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. "Intriguing Properties of Neural Networks." https://arxiv.org/abs/1312.6199.

Tank, Aytekin. 2017. "This Is the Year of the Machine Learning Revolution." Edited by Entrepreneur Magazine. January 12, 2017. https://www.entrepreneur.com/leadership/this-is-the-year-of-the-machine-learning-revolution/287324.

Teh, Yee Whye, Max Welling, Simon Osindero, and Geoffrey E. Hinton. 2003. "Energy-Based Models for Sparse Overcomplete Representations." *J. Mach. Learn. Res.* 4 (null): 1235–60.

Teney, Damien, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. "Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision." In *Computer Vision - ECCV 2020*, 580–99. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-030-58607-2_34.

Tolomei, Gabriele, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. "Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 465–74. https://doi.org/10.1145/3097983.3098039.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." https://arxiv.org/abs/2302.13971.

Trinh, T. H., Wu, Y., Le, and Q. V. et al. 2024. "Solving olympiad geometry without human demonstrations." *Nature 625*, 476–82. https://doi.org/https://doi.org/10.1038/s41586-023-06747-5.

Upadhyay, Sohini, Shalmali Joshi, and Himabindu Lakkaraju. 2021. "Towards Robust and Reliable Algorithmic Recourse." *Advances in Neural Information Processing Systems* 34: 16926–37.

Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. https://doi.org/10.1145/3287560.3287566.

Van Prooijen, Jan-Willem, Karen M Douglas, and Clara De Inocencio. 2018. "Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural." *European Journal of Social Psychology* 48 (3): 320–35.

Vardi, Moshe Y. 2018. "Vardi's Insights: Move Fast and Break Things." 2018. https://doi.org/10.1145/3244026.

Varshney, Kush R. 2022. *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published.

Venkatasubramanian, Suresh, and Mark Alfano. 2020. "The Philosophical Basis of Algorithmic Recourse." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–93. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372876.

Verma, Sahil, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. 2022. "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review." https://arxiv.org/abs/2010.10596.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harv. JL & Tech.* 31: 841. https://doi.org/10.2139/ssrn.3063289.

Walker, Alexander C, Martin Harry Turpin, Jennifer A Stolz, Jonathan A Fugelsang, and Derek J Koehler. 2019. "Finding meaning in the clouds: Illusory pattern perception predicts receptivity to pseudo-profound bullshit." *Judgment and Decision Making* 14 (2): 109–19.

Wang, Zhou, Eero P Simoncelli, and Alan C Bovik. 2003. "Multiscale Structural Similarity for Image Quality Assessment." In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2:1398–1402. Ieee.

Waytz, Adam, Nicholas Epley, and John T Cacioppo. 2010. "Social cognition unbound: Insights into anthropomorphism and dehumanization." *Current Directions in Psychological Science* 19 (1): 58–62.

Weissburg, Iain Xie, Mehir Arora, Liangming Pan, and William Yang Wang. 2024. "Tweets to Citations: Unveiling the Impact of Social Media Influencers on AI Research Visibility." *arXiv Preprint arXiv:2401.13782*.

Welling, Max, and Yee W Teh. 2011. "Bayesian Learning via Stochastic Gradient Langevin Dynamics." In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 681–88. Citeseer.

Widmer, Gerhard, and Miroslav Kubat. 1996. "Learning in the Presence of Concept Drift and Hidden Contexts." *Machine Learning* 23 (1): 69–101. https://doi.org/10.1007/bf00116900.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 1–9.

Wilson, Andrew Gordon. 2020. "The Case for Bayesian Deep Learning." https://arxiv.org/abs/2001.10995.

Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6707–23. Online: ACL. https://doi.org/10.18653/v1/2021.acl-long.523.

Xiao, Han, Kashif Rasul, and Roland Vollgraf. 2017. "Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms." https://arxiv.org/abs/1708.07747.

Yeh, I-Cheng. 2016. "Default of Credit Card Clients." UCI Machine Learning Repository.

Yeh, I-Cheng, and Che-hui Lien. 2009. "The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients." *Expert Systems with Applications* 36 (2): 2473–80. https://doi.org/10.1016/j.eswa.2007.12.020.

Zečević, Matej, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. "Causal Parrots: Large Language Models May Talk Causality but Are Not Causal." *arXiv Preprint arXiv:2308.13067*.

Zenil, Hector. 2024. "Curb The Enthusiasm." https://www.linkedin.com/posts/zenil_google-deepmind-makes-breakthrough-in-solving-activity-7154157779136446464-Gvv-.

Zgraggen, Emanuel, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. "Investigating the effect of the multiple comparisons problem in visual analysis." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Commun. ACM* 64 (3): 107–15. https://doi.org/10.1145/3446776.

CLOSING REMARKS

Around the beginning of this Ph.D. trajectory, my mother told me that apparently as a young kid I was fascinated with the 2001 movie, "A.I.", by Steven Spielberg. According to her, at just 8 years of age, I must have been way more receptive to the potential impact of artificial intelligence than herself, evidently already paving the way for what would be my future in this field. Like most parents, I think my mother may have given me a little too much credit there, because all I remember from that movie is "the kid who played young Anakin Skywalker" (or at least that is who I thought it was[5]).

Still, after initially dreaming of becoming an actor myself only to eventually get a degree in economics and work in monetary policy for a while, somehow I ended up spending the past four years of my life researching the field that gave that Spielberg movie its name. So, how exactly does a former central banker end up pursuing a Ph.D. in Trustworthy AI?

My time at the Bank of England coincided with a number of organizational transformations that were geared towards embracing technologies and data sources, that in the eyes of most economists would have been considered unconventional at the time. The Bank's Advanced Analytics (AA) division had been founded just three years before I entered the "Old Lady of Threadneedle Street"[6] for the first time in the summer of 2017, then as an intern. While many of us at the Bank marvelled at the innovative research coming out of AA, its impact on policy decisions was probably fairly described as "tangential" (at least from my perspective as an analyst who regularly contributed briefing rounds of the Bank's Monetary Policy Committee). Even though the terms "big data" and later "machine learning" would grab peoples' attention during meetings, I sensed a certain reluctance amongst colleagues and superiors to substitute tried and trusted tools for new technologies that few of us understood very well.

At the time, I was naively optimistic that the necessary understanding could be swiftly acquired, and we would soon replace all of our ordinary least squares regressions with gradient-boosting trees and universal function approximators (a.k.a. artificial neural networks). Full of enthusiasm to "start building", I went back to Barcelona School of Economics to study for a Master Degree in Data Science, and—as it

[5] As it turns out, the child actor playing the advanced robotic boy in "A.I." is Haley Joel Osment, who auditioned to play young Anakin but did not actually get the role. Young Anakin was in fact played by Jake Lloyd. The more you know …

[6] The Bank's nickname dates back to cartoon published in 1797: https://www.bankofengland.co.uk/explainers/who-is-the-old-lady-of-threadneedle-street.

turns out—to find out just how blatantly ignorant I had been about these promising new technologies. For the first time, I realized that the brave new world of machine learning clashed on fundamental levels with principles I had been taught in traditional econometrics: suddenly, we spent hours optimizing models for accuracy on Kaggle benchmark datasets with little to no attention to the underlying data itself. The assignment had completely changed: from understanding *why* things happen, to predicting *what* things happen. To be fair, we did cover explainability on the fringes even in these types of courses, but it still started to dawn on me that central banks would not, in fact, trust the Deep Vector Autoregressive Models we proposed in our master's project to produce future inflation forecasts (Agustí, Costa, and Altmeyer 2023). After all, public policymakers, and the people subjected to their decisions, do care about *why* things happen, not just *what* things happen. So, that is how I ended up pursuing this Ph.D. in Trustworthy AI, convinced—as I still am—that the trustworthiness of these models can and should be improved, even though it might not be easy.

Now at the end of this journey, I remain cautiously optimistic that we can continue to make progress towards trustworthy AI. Despite having witnessed practices and trends that concern me, I still believe it is possible to embrace and promote innovation and progress without adhering to premature paradigms like "moving fast and breaking things". This will require patience, persistence and effort—virtues that I believe are being undervalued in many places and communities of today's fast-paced world. Personally, I believe that trying to adhere to these virtues has played an important role in obtaining this Ph.D. degree, outweighed only by the importance of the many people involved in this journey.

ACKNOWLEDGEMENTS

A Ph.D. is often thought of as a somewhat lonely journey testing people's ability to perform research in an independent and self-reliant manner. While this certainly applies in some sense and I have often enjoyed the freedom to work autonomously on topics of my choice, I have relied on countless people to get to this stage of the process. In this section, I want to acknowledge and thank some of these people.

First and foremost, I want to thank my partner, Daniëlle Sophie Kotter, who has worn many hats throughout this journey including 'home office companion', 'therapist', 'interpreter' and 'travel buddy', to name just a few. Having her by my side has made this all an overwhelmingly happy journey, notwithstanding the emotional and mental challenges that were occasionally and perhaps inevitably brought on by looming paper deadlines. Next to Dani, my friends and family have played a similar role in making me feel supported both in terms of my academic and professional aspirations and also outside that. Thank you all for always being there.

On the professional side, I want to begin by thanking Cynthia, who has been my daily supervisor. It has been said before that the relationship with supervisors can make or break Ph.D. students, and I would wholeheartedly agree with that.

Cynthia has provided me with great academic freedom and autonomy, demonstrated faith in my work even when I had self-doubts, and—by setting such an admirable and impressive example herself—motivated me to approach this degree with the discipline and integrity it demands. Her scientific rigor, public outreach and genuine care for people, have always made me feel immensely privileged and proud to work under her guidance. Cynthia, thank you for everything you have done for me and everything you continue to do for the people around you! I am immensely grateful and continue to be amazed and inspired by everything you achieve. I am also sincerely grateful to Arie, who has consistently been supportive of all aspects of this project and frequently provided the necessary input to set individual projects on their right path. Arie has played an important role in ensuring that we present our work in the right manner and to the right audiences, which I am very grateful for. Especially during the early stages of this Ph.D., he has pushed me to aim higher than perhaps I would have, without inflicting unnecessary pressure. Thank you both, for always making me feel supported and appreciated, for motivating me through your interest in my work, and for creating many unique opportunities and connections.

Of course, I also want to express my gratitude to my many co-authors, colleagues and contributors. I have been asked once or twice why I always speak in the first-person plural when presenting work from my Ph.D. What by now has become force-of-habit, has always felt very natural to me: even though I feel a great sense of personal responsibility for the outcomes of this project, I believe that everyone who has been involved in one way or another also has a stake in it. I am very grateful to all of you and apologize to anyone I have missed: Jaehun Kim, Andrew Demetriou, Sandy Manolios, Marijn Roelvink, Imara van Dinten, Mojtaba Farmanbar, Elvan Kula, Leonhard Applis, Lorena Poenaru-Olaru, Floris den Hengst, Luís Cruz, Sara Salimzadeh, George Siachamis, Pasquale Caterino, Jorge Luiz Franco.

I also want to thank the many students at TU Delft I had the privilege to work with throughout this PhD. In particular, I want to highlight two exceptional students who I have not only had the privilege to supervise, but also directly collaborate with on research that is included in this thesis: Aleksander Buszydlik and Karol Dobiczek. Both have contributed as co-authors to Chapter 3 after demonstrating remarkable skill and enthusiasm during their bachelor's research projects. Two years later, Cynthia and I again had the pleasure of working with them on their master's research projects, both of which have resulted in peer-reviewed publications. Finally, Aleksander also contributed substantially to Chapter 5, once again demonstrating tireless enthusiasm and team spirit. Thank you both, I have really enjoyed working with you!

As I am finishing this thesis off for printing, I also want to express my gratitude to my Ph.D. committee and others who have provided valuable feedback and thoughts on thesis throughout: prof. dr. ir. Jan H. Kwakkel, prof. dr. Iman P.P. van Lelyveld, dr. Margaret Mitchell, prof. dr. Mykola Pechenizkiy, dr. Flavia Barsotti and dr. Jiahao Chen.

Last but definitely not least, I want to thank the open-source software community, most notably Julia and Quarto, for creating such a welcoming and supportive environment for researchers like myself. A huge proportion of my time spent on research has involved open-source software development, and both Julia and Quarto have made that time enjoyable. Julia and its community, I strongly believe, have made me a better developer and thereby enabled me to be a better researcher. Quarto has allowed me to prototype, produce and publish science in the way that I think it should be. Amongst other things, I have used Quarto to build this very thesis you are reading. I feel a deep sense of gratitude to both of these communities for making their products openly accessible.

A

PUBLICATIONS

The following is a list of publications falling into one of the following two categories:

1. The publication is included as chapter in this thesis.
2. The publication was released during Patrick's Ph.D. trajectory and: (a) lists Patrick as a (co-)author; (b) is broadly related to this thesis.

All of the papers listed here have either already been published or they have been accepted for publication, unless otherwise stated below:

- Buszydlik et al. (2024) received positive reviews at NeurIPS 2024, but ultimately ended up as a border-line 'reject' on the basis of not being going fit for the venue.

ACADEMIC RESEARCH

Patrick Altmeyer, Aleksander Buszydlik, Arie Deursen, Cynthia C. S. Liem (2026). 'Counterfactual Training: Teaching Models Plausible and Actionable Explanations'. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Available at: upcoming. (**Chapter 5**)

Aleksander Buszydlik, Patrick Altmeyer, Cynthia C. S. Liem, Roel Dobbe (2025). 'Understanding the Affordances and Constraints of Explainable AI in Safety-Critical Contexts: A Case Study in Dutch Social Welfare'. In *Electronic Government. EGOV 2025. Lecture Notes in Computer Science*. Available at: https://link.springer.com/chapter/10.1007/978-3-032-02515-9_8

Karol Dobiczek, Patrick Altmeyer, Cynthia CS Liem (2025). 'Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics'. In *Proceedings of the Fourth Workshop on*

A

Generation, Evaluation and Metrics (GEM^2), 958–972. Available at: https://aclanthology.org/2025.gem-1.75.pdf

Aleksander Buszydlik, Patrick Altmeyer, Cynthia C. S. Liem, Roel Dobbe (2024). 'Grounding and Validation of Algorithmic Recourse in Real-World Contexts: A Systematized Literature Review'. Available at: https://openreview.net/pdf?id=oEmyoy5H5P

Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, Cynthia C. S. Liem (2024). 'Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals'. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 10829–10837, (38). DOI: https://doi.org/10.1609/aaai.v38i10.28956. (**Chapter 4**)

Patrick Altmeyer, Andrew M Demetriou, Antony Bartlett, Cynthia C. S. Liem (2024). 'Position: Stop Making Unscientific AGI Performance Claims'. In *International Conference on Machine Learning*, 1222–1242. Available at: https://proceedings.mlr.press/v235/altmeyer24a.html. (**Chapter 6**)

Floris Hengst, Ralf Wolter, Patrick Altmeyer, Arda Kaygan (2024). 'Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition'. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2412–2432. DOI: https://doi.org/10.18653/v1/2024.findings-naacl.156

Marc Agustí, Ignacio Vidal-Quadras Costa, Patrick Altmeyer (2023). 'Deep vector autoregression for macroeconomic data'. In *IFC Bulletins chapters*, (59). Available at: https://www.bis.org/ifc/publ/ifcb59_39.pdf

Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, Cynthia C. S. Liem (2023). 'Endogenous Macrodynamics in Algorithmic Recourse'. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 418–431. DOI: https://doi.org/10.1109/satml54575.2023.00036. (**Chapter 3**)

Patrick Altmeyer, Arie van Deursen, Cynthia C. S. Liem (2023). 'Explaining Black-Box Models through Counterfactuals'. In *Proceedings of the JuliaCon Conferences*, 130, (1). DOI: https://doi.org/10.21105/jcon.00130. (**Chapter 2**)

RESEARCH SOFTWARE

Patrick Altmeyer, contributors (2025). 'CounterfactualExplanations.jl'. DOI: https://doi.org/10.5281/zenodo.8239378

Patrick Altmeyer, contributors (2024). 'ConformalPrediction.jl'. DOI: https://doi.org/10.5281/zenodo.12799930

Patrick Altmeyer, contributors (2024). 'LaplaceRedux.jl'. DOI: https://doi.org/10.5281/zenodo.13758044

B

SUPERVISION

The author of this dissertation supervised multiple bachelor's and master's students at TU Delft. He also mentored external students looking to contribute to open-source software as part of Julia Seasons of Contributions and Google Summer of Code.

B.1. MASTER'S STUDENTS

Aleksander Buszydlik (2024). 'Finding Recourse for Algorithmic Recourse'. Available at: https://resolver.tudelft.nl/uuid:be47ad5a-5a4b-457c-b214-35c6c78cae36

Karol Dobiczek (2024). 'Natural Language Counterfactual Explanations in Financial Text Classification'. Available at: https://resolver.tudelft.nl/uuid:66730110-d296-4a57-b382-e9a6cc0a4aa5

Ivor Zagorac (2024). 'A Study on Counterfactual Explanations'. Available at: https://resolver.tudelft.nl/uuid:6e2c240c-03c6-4e0e-af2c-5d257e77c77c

Marit E. Radder (2024). 'A counterfactual-based evaluation framework for machine learning models that use gene expression data'. Available at: https://resolver.tudelft.nl/uuid:4cf92f8f-2a4c-43e8-9746-2ff33ca65de5

B.2. BACHELOR'S STUDENTS

Dimitar Nikolov (2024). 'How Does Predictive Uncertainty Quantification Correlate with the Plausibility of Counterfactual Explanations'. Available at: https://resolver.tudelft.nl/uuid:b0ecc3fe-4454-4c44-a624-5d335d108634

Rithik Appachi Senthilkumar (2024). 'Are Neural Networks Robust to Gradient-Based Adversaries Also More Explainable? Evidence from Counterfactuals'. Available at: https://resolver.tudelft.nl/uuid:47786bb4-ae24-4972-94a0-1bd18d756486

Giacomo Pezzali (2024). 'Do Joint Energy-Based Models Produce More Plausible Counterfactual Explanations?'. Available at: https://resolver.tudelft.nl/uuid:afe2d50d-f4b3-403f-b0e7-a0b8ede96bb0

Ali Faruk Yücel (2024). 'Metrics to Ascertain the Plausibility and Faithfulness of Counterfactual Explanations'. Available at: https://resolver.tudelft.nl/uuid:d80b688c-b0f6-4c88-a0a2-891d738f25d4

Ipek Iscan (2024). 'Advancing Explainability in Black-Box Models'. Available at: https://resolver.tudelft.nl/uuid:e50c1cae-d579-405a-9089-86a0ca925086

Giovan Angela (2023). 'Endogenous Macrodynamics in Algorithmic Recourse'. Available at: https://resolver.tudelft.nl/uuid:5023154a-53c6-44ca-9d09-1670ba0ded31

Aleksander Buszydlik (2022). 'Quantifying the Endogenous Domain and Model Shifts Induced by the DiCE Generator'. Available at: https://resolver.tudelft.nl/uuid:cb0bf4ac-4055-489b-b768-e5b53ec6fa47

Karol Dobiczek (2022). 'Quantifying the Endogenous Domain and Model Shifts Induced by the CLUE Recourse Generator'. Available at: https://resolver.tudelft.nl/uuid:6a249d72-9e1e-4e81-abdc-463260c7d1bc

B.3. EXTERNAL

Google Summer of Code 2024
Together with Mojtaba Farmanbar, the author of this dissertation mentored Pasquale Caterino, who contributed support for conformalized Bayes to our ConformalPrediction.jl (Caterino 2024).
Julia Season of Contributions 2024
Together with Moritz Schauer, Patrick mentored Jorge Luiz Franco, who contributed support for causal algorithmic recourse to our CounterfactualExplanations.jl (Luiz Franco 2024).

Jorge Luiz Franco (2024). 'JSoC: When Causality Meets Recourse'. Available at: https://www.taija.org/blog/posts/causal-recourse/

Pasquale Caterino (2024). 'Google Summer of Code 2024 Final Report: Add support for Conformalized Bayes to ConformalPrediction.jl'. Available at: https://gist.github.com/pasq-cat/f25eebc492366fb6a4f428426f93f45f

C

CURRICULUM VITAE

Patrick Altmeyer was born on March 28, 1993 in Düsseldorf, Germany. From 2003 until 2012, he followed his secondary school education at the Geschwister-Scholl-Gymnasium, Düsseldorf, completing the Abitur with one of the highest final grades in the state of North-Rhine-Westphalia. Upon graduation, he was the recipient of the prestigious "Deutschlandstipendium" scholarship and recognized by the Deutsche Physikalische Gesellschaft for achieving outstanding scores in Physics.

After this, he pursued the Degree of Master of Arts (2013-2017) at the University of Edinburgh, which he completed with First Class Honours in Economics as one of the top three students of his cohort. He received the School of Economics Prize for academic excellence in Economics (2013) and the Joint Prize for the best performance in Economics (2017). During these studies, he also spent one year studying abroad at Universitat Pompeu Fabra, held a TEDx talk on European integration and gained professional experience through multiple internships at the Handelsblatt Research Institute and part-time teaching activities for the School of Economics.

Following his undergraduate degree, Patrick Altmeyer worked as post-graduate intern for the Bank of England before pursuing a postgraduate degree at Barcelona School of Economics (2017-2018). The institution awarded him a Full-Tuition Scholarship and he obtained the Master Degree in Economics and Finance, achieving the highest overall grade in the finance program. He then returned to the Bank of England as a graduate economist (2018-2021), where he was involved in research, market intelligence, policy briefing work and software development. He presented research on the demand for central bank reserves at the European Central Bank (2019) and co-authored a staff working paper on yield curve sensitivity (2023). During his final year of employment, the Bank of England funded his pursuit of the Master Degree in Data Science at Barcelona School of Economics (2020-2021), which he completed with excellent grades.

From 2021, Patrick Altmeyer pursued the Ph.D. degree at Delft University of Technology under the supervision of Dr. ir. Cynthia C. S. Liem and Prof. dr. ir. Arie van Deursen. The program was funded by AI for Fintech Research—a five-year collaboration between ING and Delft University of Technology and a participating lab of

the Innovation Center for Artificial Intelligence (ICAI). His research collaborations with ING colleagues Mojtaba Farmanbar, Flavia Barsotti and Floris den Hengst led to publications in top-tier venues (Altmeyer, Farmanbar, et al. 2024a; Hengst et al. 2024) and a 1st Prize Win at ING Experiment Week 2023. He also published research at the First IEEE Conference on Secure and Trustworthy Machine Learning (2023) and the Forty-First International Conference on Machine Learning (2024).

During his Ph.D. studies, Patrick Altmeyer gained further industry exposure through invited talks at the Bank of England, De Nederlandsche Bank, The Alan Turing Institute and the TÜV AI.Lab, among others. He also proactively reached out to the public by maintaining a blog[1] focused on communicating his research in an accessible manner. Several of his blog post were featured as editor's picks on the popular online publication, Towards Data Science. He also founded and maintained Taija[2], an open-source software organization for Trustworthy AI in Julia. Taija has earned great recognition in the Julia community with multiple presentations at JuliaCon Global, Google Summer of Code projects and a total of more than 300 stars on GitHub. Among the list of contributors are colleagues from ING, as well as former bachelor's and master's students from Delft University of Technology who Patrick Altmeyer (co-)supervised throughout his Ph.D. Outside of the professional realm, Patrick Altmeyer combined his passion for sports with a charitable cause in 2024, when he organized a fundraiser for Mental Health Europe that received more than €1,000 in donations.

Following his Ph.D. defense in early 2026, Patrick Altmeyer will be on the job market. He is open to roles in both academia and industry.

[1] https://www.patalt.org/blog/
[2] https://www.taija.org/

D

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

D.1. DETAILED RESULTS: SYNTHETIC DATA

D.1.1. LINE CHARTS

The evolution of the evaluation metrics over the course of the experiment is shown for different datasets in Figure D.1 to Figure D.4.

D.1.2. ERROR BAR CHARTS

The evaluation metrics at the end of the experiment are shown for different datasets in Figure D.5 to Figure D.8.

D.1.3. STATISTICAL SIGNIFICANCE

Table D.1 presents the tests for statistical significance of the estimated MMD metrics.

Table D.1. Tests for statistical significance of the estimated MMD metrics. We have highlighted p-values smaller than the significance level $\alpha = 0.05$ in bold. Data: Synthetic.

Metric	Data	Generator	Model	p-value
MMD	Circles	DICE	Deep Ensemble	0.988
MMD	Circles	DICE	Linear	1.0
MMD	Circles	DICE	MLP	0.99
MMD	Circles	Generic (=0.5)	Deep Ensemble	0.996
MMD	Circles	Generic (=0.5)	Linear	0.996

Continued below.

D

Metric	Data	Generator	Model	p-value
MMD	Circles	Generic (=0.5)	MLP	0.99
MMD	Circles	Greedy	Deep Ensemble	0.992
MMD	Circles	Greedy	Linear	1.0
MMD	Circles	Greedy	MLP	0.994
MMD	Circles	Latent	Deep Ensemble	0.9975
MMD	Circles	Latent	Linear	0.9925
MMD	Circles	Latent	MLP	1.0
MMD	Linearly Separable	DICE	Deep Ensemble	**0.0**
MMD	Linearly Separable	DICE	Linear	**0.0**
MMD	Linearly Separable	DICE	MLP	**0.0**
MMD	Linearly Separable	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Linearly Separable	Generic (=0.5)	Linear	**0.0**
MMD	Linearly Separable	Generic (=0.5)	MLP	**0.0**
MMD	Linearly Separable	Greedy	Deep Ensemble	**0.0**
MMD	Linearly Separable	Greedy	Linear	**0.0**
MMD	Linearly Separable	Greedy	MLP	**0.0**
MMD	Linearly Separable	Latent	Deep Ensemble	0.748
MMD	Linearly Separable	Latent	Linear	0.768
MMD	Linearly Separable	Latent	MLP	0.69
MMD	Moons	DICE	Deep Ensemble	**0.0**
MMD	Moons	DICE	Linear	**0.0**
MMD	Moons	DICE	MLP	**0.0**
MMD	Moons	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Moons	Generic (=0.5)	Linear	**0.0**
MMD	Moons	Generic (=0.5)	MLP	**0.0**
MMD	Moons	Greedy	Deep Ensemble	**0.0**
MMD	Moons	Greedy	Linear	**0.0**
MMD	Moons	Greedy	MLP	**0.0**
MMD	Moons	Latent	Deep Ensemble	**0.0**
MMD	Moons	Latent	Linear	**0.0**
MMD	Moons	Latent	MLP	**0.0**
MMD	Overlapping	DICE	Deep Ensemble	**0.0**
MMD	Overlapping	DICE	Linear	**0.0**
MMD	Overlapping	DICE	MLP	**0.0**
MMD	Overlapping	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Overlapping	Generic (=0.5)	Linear	**0.0**
MMD	Overlapping	Generic (=0.5)	MLP	**0.0**
MMD	Overlapping	Greedy	Deep Ensemble	**0.0**
MMD	Overlapping	Greedy	Linear	**0.0**
MMD	Overlapping	Greedy	MLP	**0.0**
MMD	Overlapping	Latent	Deep Ensemble	**0.0**
MMD	Overlapping	Latent	Linear	**0.0**
MMD	Overlapping	Latent	MLP	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD	Circles	DICE	Deep Ensemble	0.996
PP MMD	Circles	DICE	Linear	0.796
PP MMD	Circles	DICE	MLP	0.9975
PP MMD	Circles	Generic (=0.5)	Deep Ensemble	1.0
PP MMD	Circles	Generic (=0.5)	Linear	0.996
PP MMD	Circles	Generic (=0.5)	MLP	0.992
PP MMD	Circles	Greedy	Deep Ensemble	1.0
PP MMD	Circles	Greedy	Linear	**0.0**
PP MMD	Circles	Greedy	MLP	0.996
PP MMD	Circles	Latent	Deep Ensemble	0.9975
PP MMD	Circles	Latent	Linear	**0.0**
PP MMD	Circles	Latent	MLP	0.994
PP MMD	Linearly Separable	DICE	Deep Ensemble	0.9525
PP MMD	Linearly Separable	DICE	Linear	**0.0**
PP MMD	Linearly Separable	DICE	MLP	0.964
PP MMD	Linearly Separable	Generic (=0.5)	Deep Ensemble	0.958
PP MMD	Linearly Separable	Generic (=0.5)	Linear	**0.0**
PP MMD	Linearly Separable	Generic (=0.5)	MLP	0.944
PP MMD	Linearly Separable	Greedy	Deep Ensemble	0.716
PP MMD	Linearly Separable	Greedy	Linear	**0.0**
PP MMD	Linearly Separable	Greedy	MLP	0.684
PP MMD	Linearly Separable	Latent	Deep Ensemble	0.856
PP MMD	Linearly Separable	Latent	Linear	0.46
PP MMD	Linearly Separable	Latent	MLP	0.852
PP MMD	Moons	DICE	Deep Ensemble	0.865
PP MMD	Moons	DICE	Linear	**0.0**
PP MMD	Moons	DICE	MLP	0.87
PP MMD	Moons	Generic (=0.5)	Deep Ensemble	0.678
PP MMD	Moons	Generic (=0.5)	Linear	**0.0**
PP MMD	Moons	Generic (=0.5)	MLP	0.84
PP MMD	Moons	Greedy	Deep Ensemble	0.388
PP MMD	Moons	Greedy	Linear	**0.0**
PP MMD	Moons	Greedy	MLP	0.346
PP MMD	Moons	Latent	Deep Ensemble	0.902
PP MMD	Moons	Latent	Linear	**0.004**
PP MMD	Moons	Latent	MLP	0.91
PP MMD	Overlapping	DICE	Deep Ensemble	**0.0**
PP MMD	Overlapping	DICE	Linear	**0.0**
PP MMD	Overlapping	DICE	MLP	**0.002**
PP MMD	Overlapping	Generic (=0.5)	Deep Ensemble	**0.004**
PP MMD	Overlapping	Generic (=0.5)	Linear	**0.0**
PP MMD	Overlapping	Generic (=0.5)	MLP	**0.002**
PP MMD	Overlapping	Greedy	Deep Ensemble	**0.002**

Continued below.

D

D

Metric	Data	Generator	Model	p-value
PP MMD	Overlapping	Greedy	Linear	**0.0**
PP MMD	Overlapping	Greedy	MLP	**0.004**
PP MMD	Overlapping	Latent	Deep Ensemble	**0.034**
PP MMD	Overlapping	Latent	Linear	**0.012**
PP MMD	Overlapping	Latent	MLP	**0.034**
PP MMD (grid)	Circles	DICE	Deep Ensemble	0.762
PP MMD (grid)	Circles	DICE	Linear	0.814
PP MMD (grid)	Circles	DICE	MLP	0.7375
PP MMD (grid)	Circles	Generic (=0.5)	Deep Ensemble	0.89
PP MMD (grid)	Circles	Generic (=0.5)	Linear	0.994
PP MMD (grid)	Circles	Generic (=0.5)	MLP	0.688
PP MMD (grid)	Circles	Greedy	Deep Ensemble	0.568
PP MMD (grid)	Circles	Greedy	Linear	**0.0**
PP MMD (grid)	Circles	Greedy	MLP	0.776
PP MMD (grid)	Circles	Latent	Deep Ensemble	1.0
PP MMD (grid)	Circles	Latent	Linear	**0.0**
PP MMD (grid)	Circles	Latent	MLP	0.996
PP MMD (grid)	Linearly Separable	DICE	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	DICE	Linear	**0.0**
PP MMD (grid)	Linearly Separable	DICE	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Greedy	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Greedy	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Greedy	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Latent	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Latent	MLP	**0.0**
PP MMD (grid)	Moons	DICE	Deep Ensemble	0.1225
PP MMD (grid)	Moons	DICE	Linear	**0.0**
PP MMD (grid)	Moons	DICE	MLP	**0.01**
PP MMD (grid)	Moons	Generic (=0.5)	Deep Ensemble	**0.016**
PP MMD (grid)	Moons	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Moons	Generic (=0.5)	MLP	**0.02**
PP MMD (grid)	Moons	Greedy	Deep Ensemble	**0.006**
PP MMD (grid)	Moons	Greedy	Linear	**0.0**
PP MMD (grid)	Moons	Greedy	MLP	**0.0**
PP MMD (grid)	Moons	Latent	Deep Ensemble	0.114
PP MMD (grid)	Moons	Latent	Linear	**0.004**
PP MMD (grid)	Moons	Latent	MLP	0.174
PP MMD (grid)	Overlapping	DICE	Deep Ensemble	**0.002**
PP MMD (grid)	Overlapping	DICE	Linear	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD (grid)	Overlapping	DICE	MLP	**0.0**
PP MMD (grid)	Overlapping	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Overlapping	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Overlapping	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Overlapping	Greedy	Deep Ensemble	**0.0**
PP MMD (grid)	Overlapping	Greedy	Linear	**0.0**
PP MMD (grid)	Overlapping	Greedy	MLP	**0.002**
PP MMD (grid)	Overlapping	Latent	Deep Ensemble	0.208
PP MMD (grid)	Overlapping	Latent	Linear	**0.02**
PP MMD (grid)	Overlapping	Latent	MLP	0.342

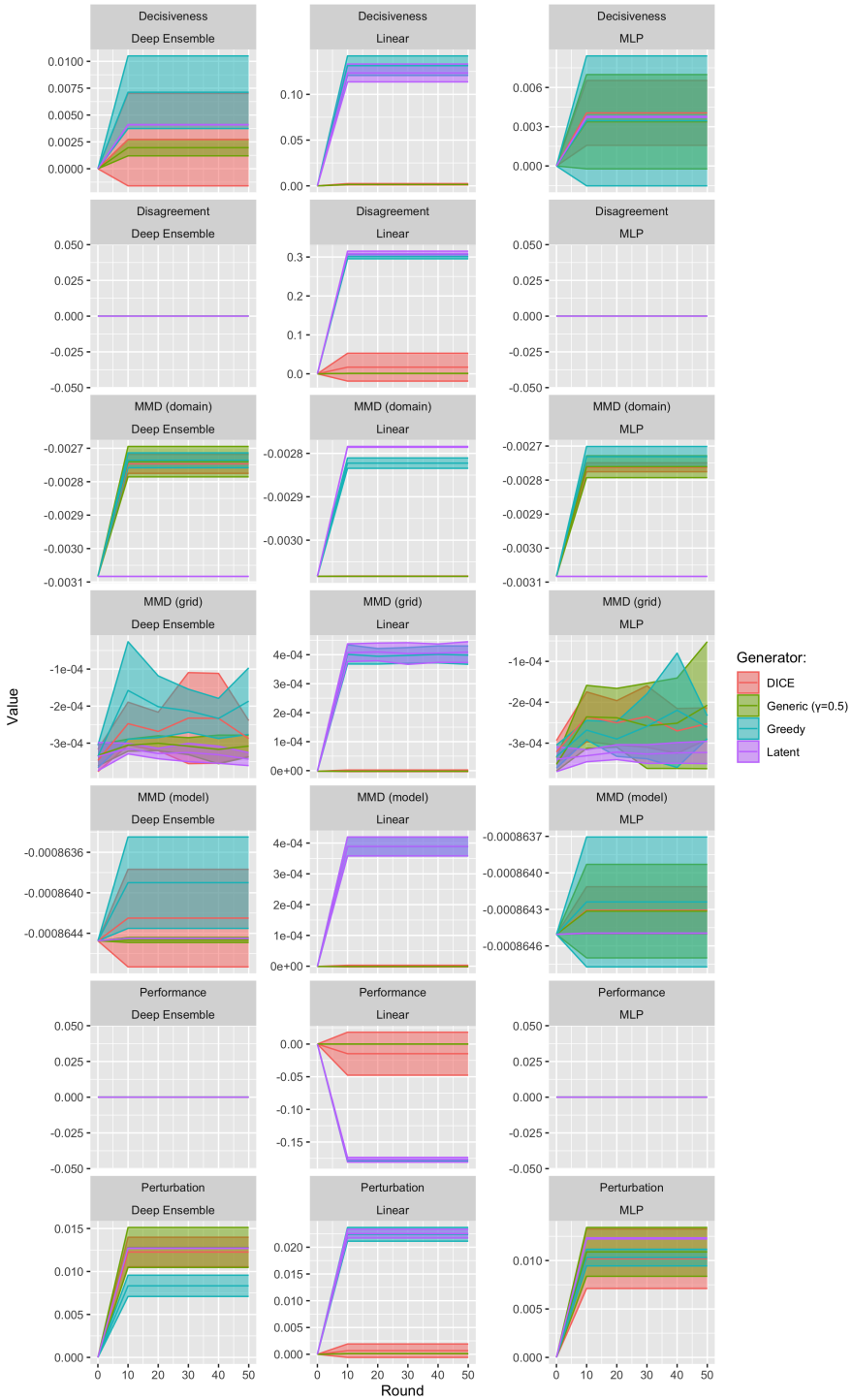


Figure D.1. Evolution of evaluation metrics over the course of the experiment. Data: Circles.

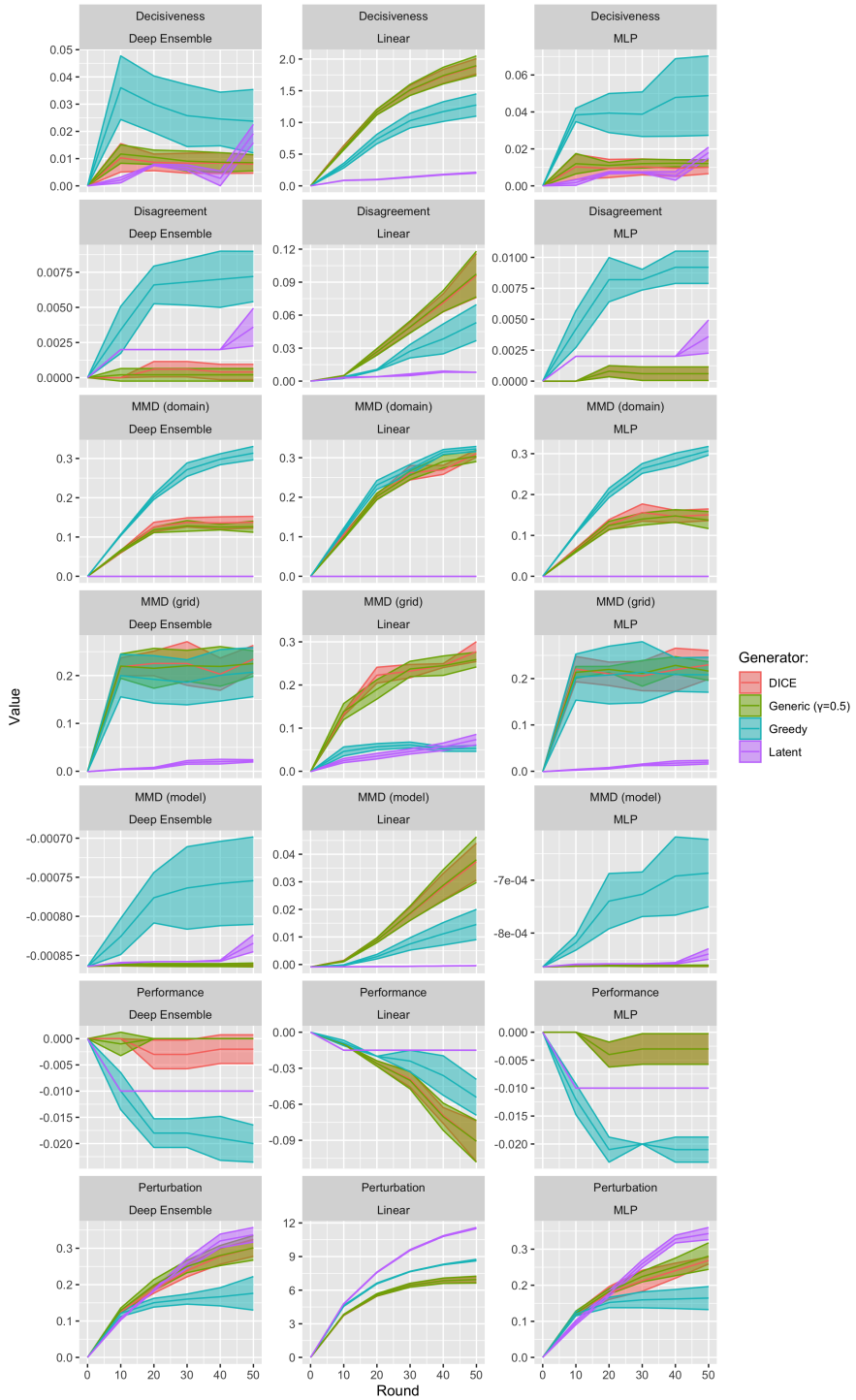


Figure D.2. Evolution of evaluation metrics over the course of the experiment. Data: Linearly Separable.

D

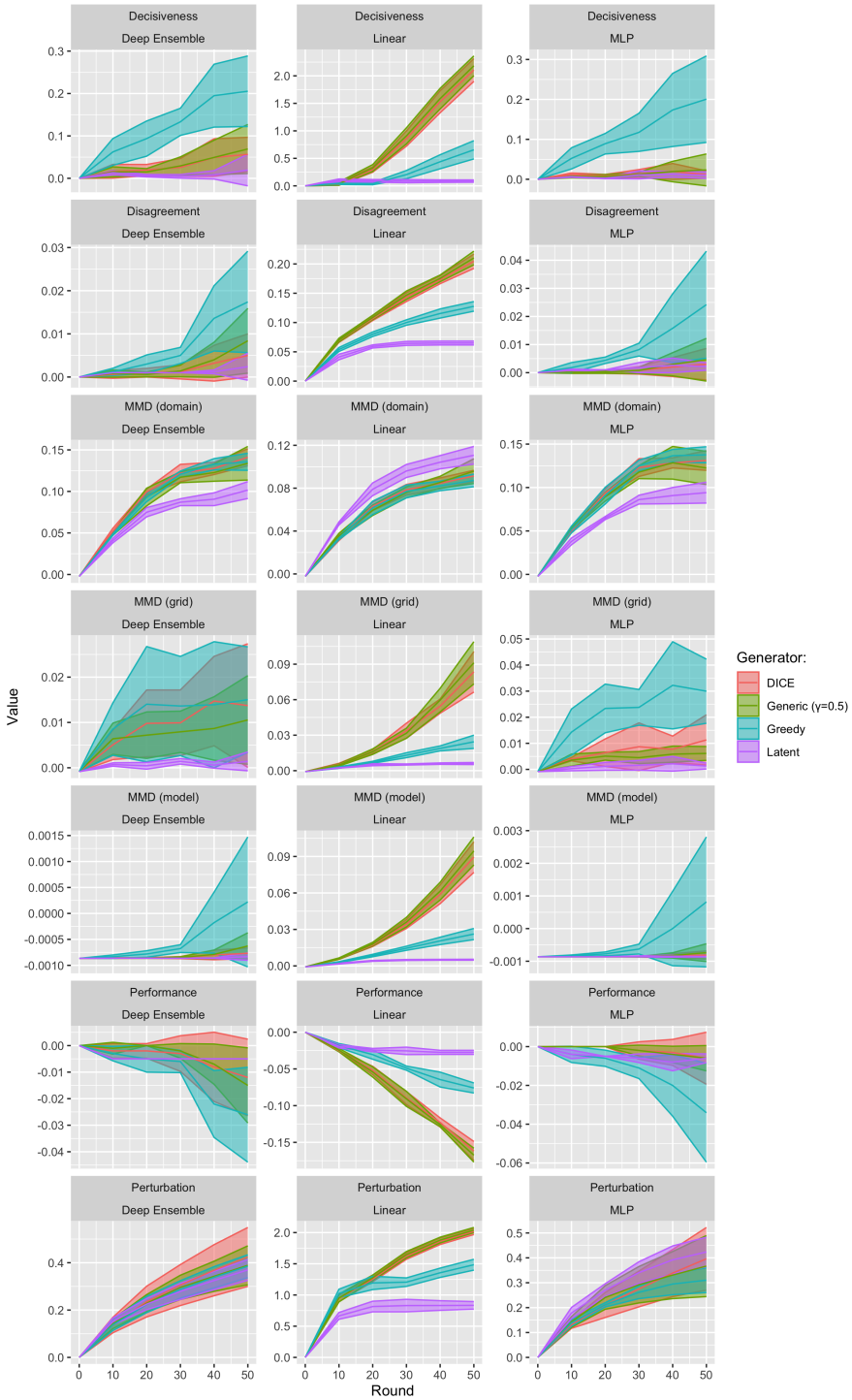


Figure D.3. Evolution of evaluation metrics over the course of the experiment. Data: Moons.

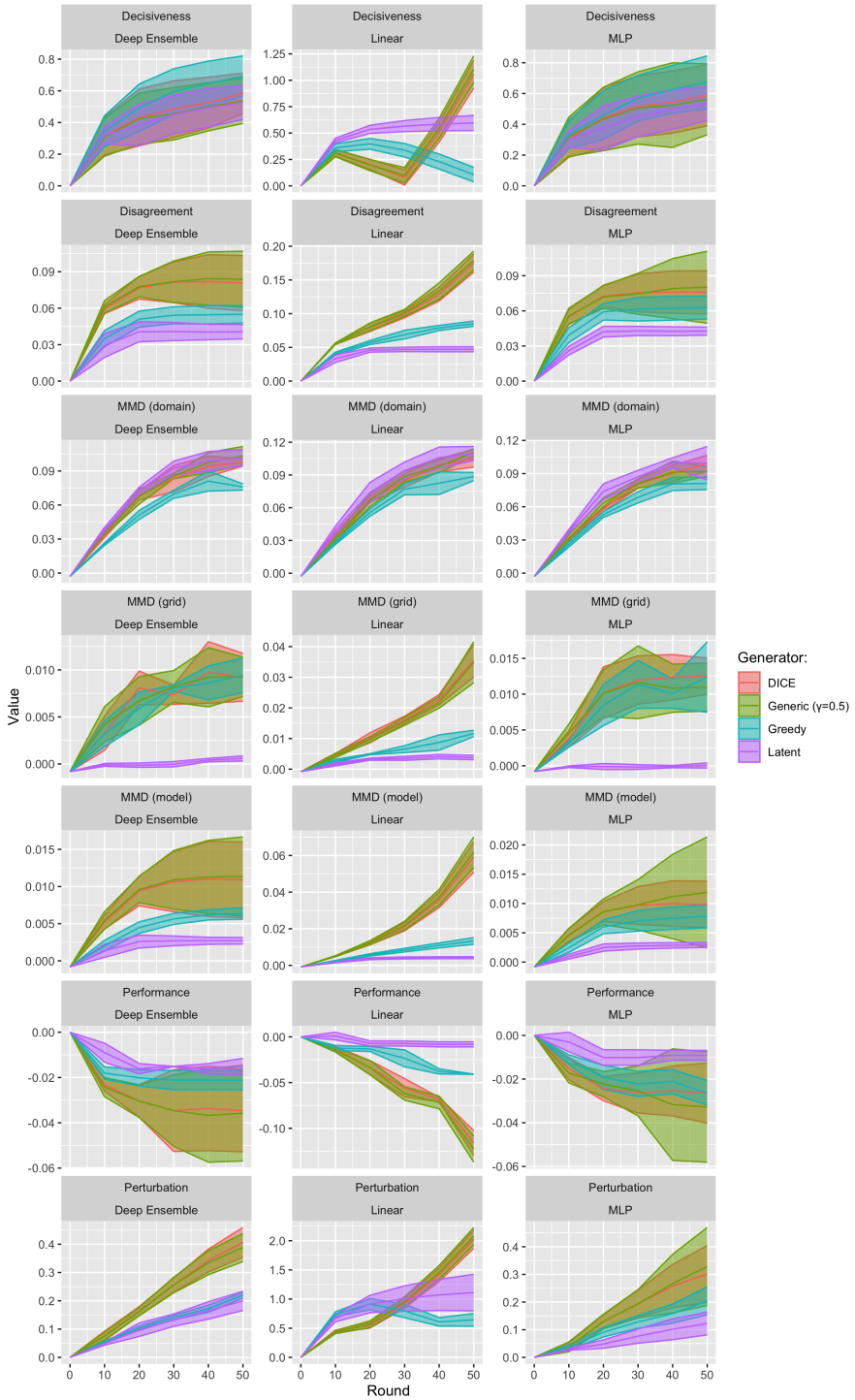


Figure D.4. Evolution of evaluation metrics over the course of the experiment. Data: Overlapping.

D

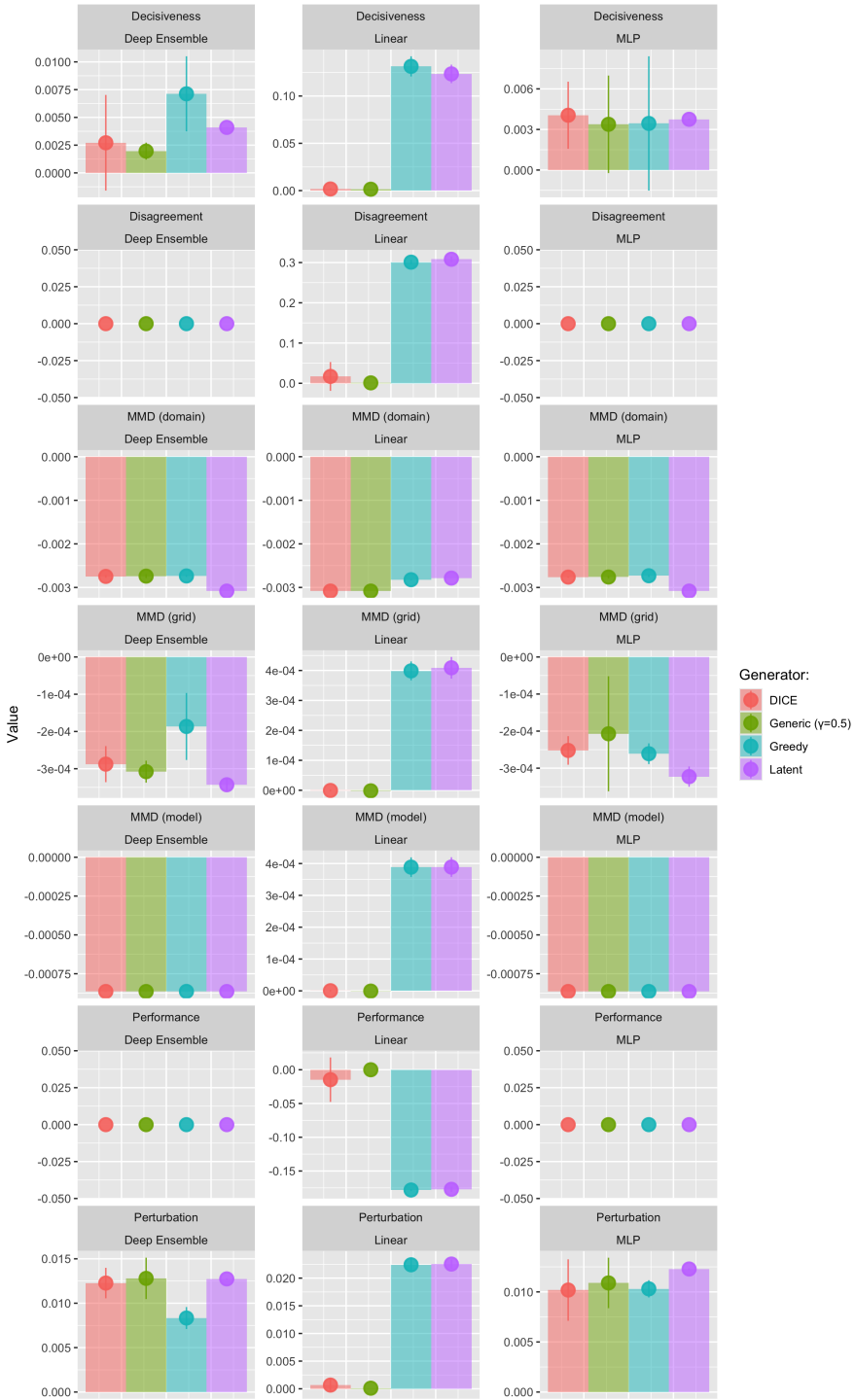


Figure D.5. Evaluation metrics at the end of the experiment. Data: Circles.

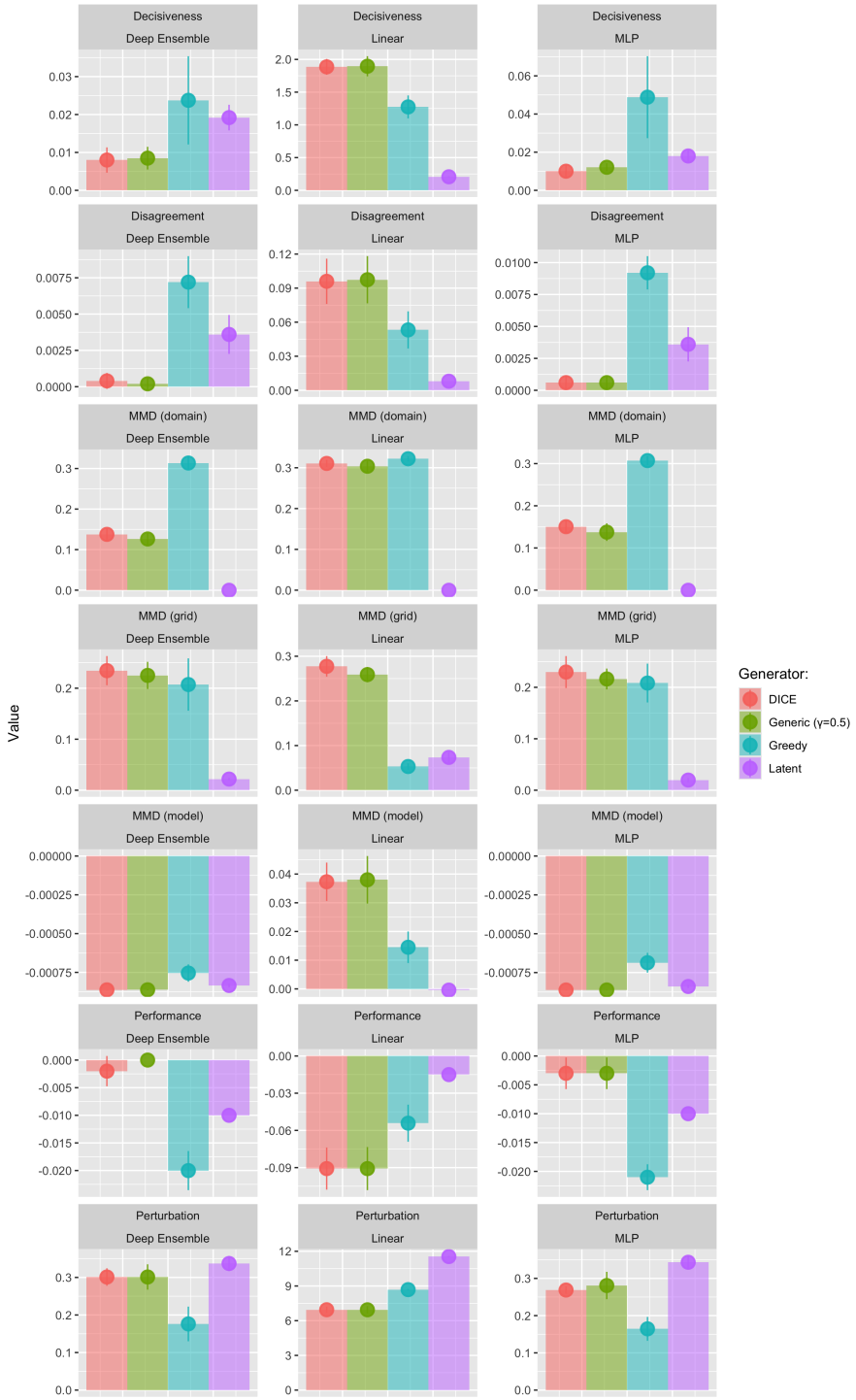


Figure D.6. Evaluation metrics at the end of the experiment. Data: Linearly Separable.

D

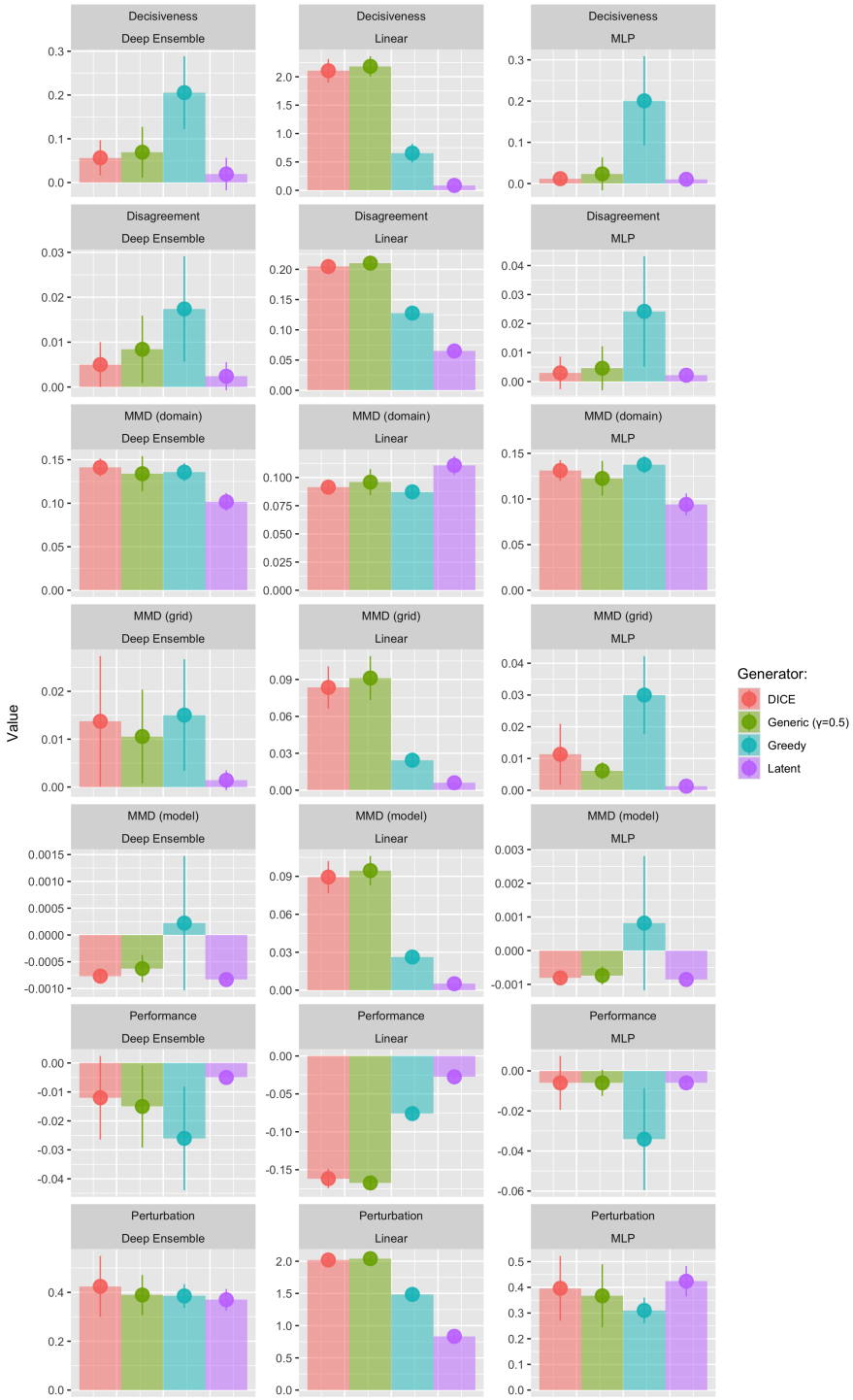


Figure D.7. Evaluation metrics at the end of the experiment. Data: Moons.

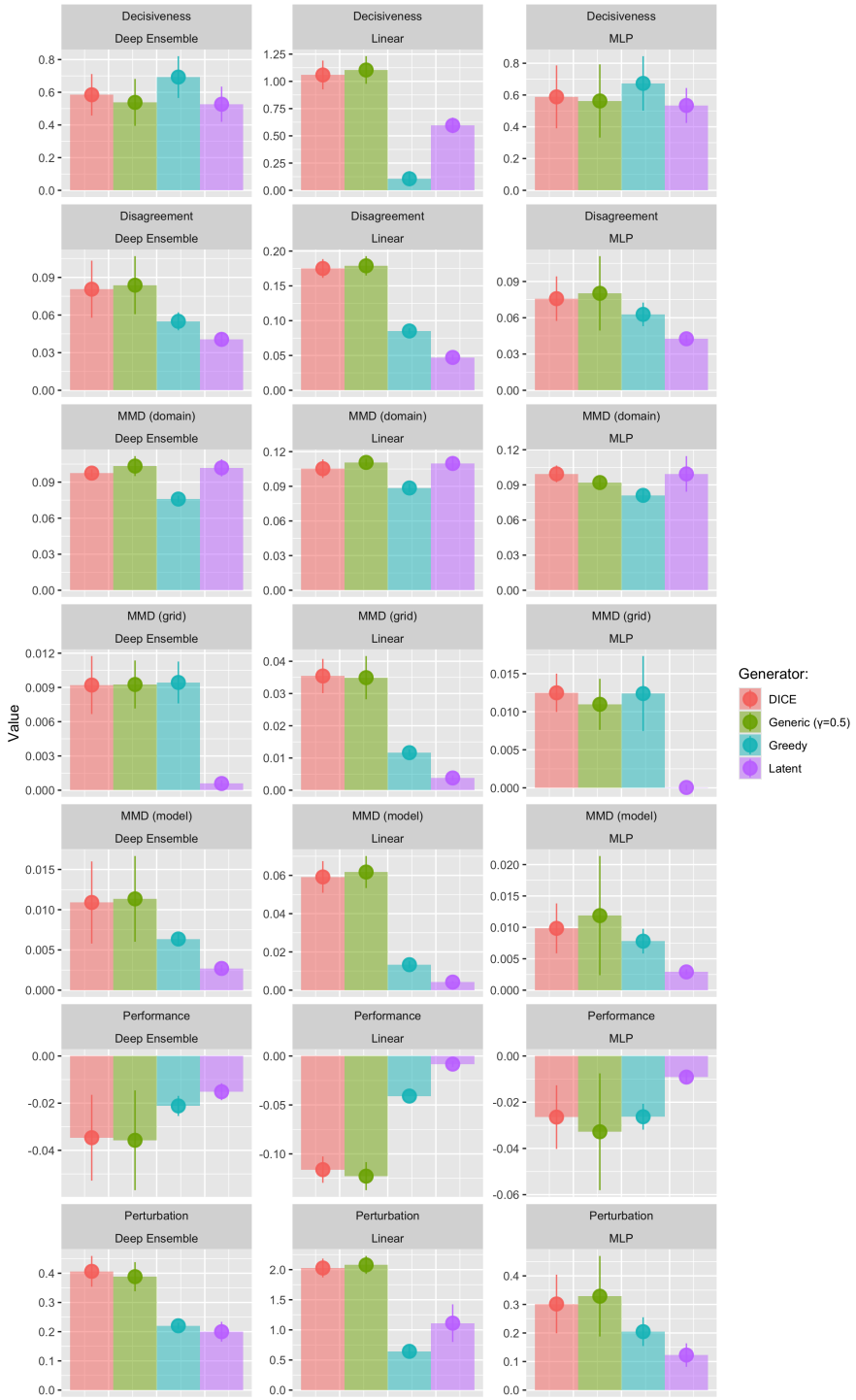


Figure D.8. Evaluation metrics at the end of the experiment. Data: Overlapping.

D

D.2. DETAILED RESULTS: REAL-WORLD DATA

D.2.1. LINE CHARTS

The evolution of the evaluation metrics over the course of the experiment is shown for different datasets in Figure D.9 to Figure D.11.

D.2.2. ERROR BAR CHARTS

The evaluation metrics at the end of the experiment are shown for different datasets in Figure D.12 to Figure D.14.

D

D.2.3. STATISTICAL SIGNIFICANCE

Table D.2 presents the tests for statistical significance of the estimated MMD metrics.

Table D.2. Tests for statistical significance of the estimated MMD metrics. We have highlighted p-values smaller than the significance level $\alpha = 0.05$ in bold. Data: Real-World.

Metric	Data	Generator	Model	p-value
MMD	Cal Housing	DICE	Deep Ensemble	**0.0**
MMD	Cal Housing	DICE	Linear	**0.0**
MMD	Cal Housing	DICE	MLP	**0.0**
MMD	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Cal Housing	Generic (=0.5)	Linear	**0.0**
MMD	Cal Housing	Generic (=0.5)	MLP	**0.0**
MMD	Cal Housing	Greedy	Deep Ensemble	**0.0**
MMD	Cal Housing	Greedy	Linear	**0.0**
MMD	Cal Housing	Greedy	MLP	**0.0**
MMD	Cal Housing	Latent	Deep Ensemble	**0.0**
MMD	Cal Housing	Latent	Linear	**0.0**
MMD	Cal Housing	Latent	MLP	**0.0**
MMD	Credit Default	DICE	Deep Ensemble	1.0
MMD	Credit Default	DICE	Linear	1.0
MMD	Credit Default	DICE	MLP	1.0
MMD	Credit Default	Generic (=0.5)	Deep Ensemble	1.0
MMD	Credit Default	Generic (=0.5)	Linear	1.0
MMD	Credit Default	Generic (=0.5)	MLP	1.0
MMD	Credit Default	Greedy	Deep Ensemble	1.0
MMD	Credit Default	Greedy	Linear	1.0
MMD	Credit Default	Greedy	MLP	1.0
MMD	Credit Default	Latent	Deep Ensemble	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
MMD	Credit Default	Latent	Linear	1.0
MMD	Credit Default	Latent	MLP	**0.0**
MMD	GMSC	DICE	Deep Ensemble	0.082
MMD	GMSC	DICE	Linear	0.51
MMD	GMSC	DICE	MLP	0.338
MMD	GMSC	Generic (=0.5)	Deep Ensemble	0.306
MMD	GMSC	Generic (=0.5)	Linear	0.278
MMD	GMSC	Generic (=0.5)	MLP	0.128
MMD	GMSC	Greedy	Deep Ensemble	**0.032**
MMD	GMSC	Greedy	Linear	**0.006**
MMD	GMSC	Greedy	MLP	**0.0**
MMD	GMSC	Latent	Deep Ensemble	**0.0**
MMD	GMSC	Latent	Linear	**0.0**
MMD	GMSC	Latent	MLP	**0.0**
PP MMD	Cal Housing	DICE	Deep Ensemble	**0.0**
PP MMD	Cal Housing	DICE	Linear	**0.0**
PP MMD	Cal Housing	DICE	MLP	**0.0**
PP MMD	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Generic (=0.5)	Linear	**0.0**
PP MMD	Cal Housing	Generic (=0.5)	MLP	**0.0**
PP MMD	Cal Housing	Greedy	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Greedy	Linear	**0.0**
PP MMD	Cal Housing	Greedy	MLP	**0.0**
PP MMD	Cal Housing	Latent	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Latent	Linear	**0.0**
PP MMD	Cal Housing	Latent	MLP	**0.0**
PP MMD	Credit Default	DICE	Deep Ensemble	**0.0**
PP MMD	Credit Default	DICE	Linear	**0.0**
PP MMD	Credit Default	DICE	MLP	**0.0**
PP MMD	Credit Default	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	Credit Default	Generic (=0.5)	Linear	**0.0**
PP MMD	Credit Default	Generic (=0.5)	MLP	**0.0**
PP MMD	Credit Default	Greedy	Deep Ensemble	**0.0**
PP MMD	Credit Default	Greedy	Linear	**0.044**
PP MMD	Credit Default	Greedy	MLP	**0.0**
PP MMD	Credit Default	Latent	Deep Ensemble	**0.0**
PP MMD	Credit Default	Latent	Linear	0.436
PP MMD	Credit Default	Latent	MLP	**0.0**
PP MMD	GMSC	DICE	Deep Ensemble	**0.032**
PP MMD	GMSC	DICE	Linear	**0.0**
PP MMD	GMSC	DICE	MLP	**0.0**
PP MMD	GMSC	Generic (=0.5)	Deep Ensemble	**0.018**
PP MMD	GMSC	Generic (=0.5)	Linear	**0.0**

Continued below.

D

Metric	Data	Generator	Model	p-value
PP MMD	GMSC	Generic (=0.5)	MLP	**0.0**
PP MMD	GMSC	Greedy	Deep Ensemble	**0.02**
PP MMD	GMSC	Greedy	Linear	**0.0**
PP MMD	GMSC	Greedy	MLP	**0.0**
PP MMD	GMSC	Latent	Deep Ensemble	**0.008**
PP MMD	GMSC	Latent	Linear	**0.0**
PP MMD	GMSC	Latent	MLP	**0.0**
PP MMD (grid)	Cal Housing	DICE	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	DICE	Linear	**0.0**
PP MMD (grid)	Cal Housing	DICE	MLP	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.5)	MLP	**0.004**
PP MMD (grid)	Cal Housing	Greedy	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	Greedy	Linear	**0.0**
PP MMD (grid)	Cal Housing	Greedy	MLP	**0.0**
PP MMD (grid)	Cal Housing	Latent	Deep Ensemble	**0.006**
PP MMD (grid)	Cal Housing	Latent	Linear	**0.01**
PP MMD (grid)	Cal Housing	Latent	MLP	**0.026**
PP MMD (grid)	Credit Default	DICE	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	DICE	Linear	**0.0**
PP MMD (grid)	Credit Default	DICE	MLP	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Credit Default	Greedy	Deep Ensemble	0.164
PP MMD (grid)	Credit Default	Greedy	Linear	**0.0**
PP MMD (grid)	Credit Default	Greedy	MLP	**0.0**
PP MMD (grid)	Credit Default	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Latent	Linear	**0.044**
PP MMD (grid)	Credit Default	Latent	MLP	**0.0**
PP MMD (grid)	GMSC	DICE	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	DICE	Linear	**0.0**
PP MMD (grid)	GMSC	DICE	MLP	**0.004**
PP MMD (grid)	GMSC	Generic (=0.5)	Deep Ensemble	**0.002**
PP MMD (grid)	GMSC	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	GMSC	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	GMSC	Greedy	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Greedy	Linear	**0.0**
PP MMD (grid)	GMSC	Greedy	MLP	**0.0**
PP MMD (grid)	GMSC	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Latent	Linear	**0.0**
PP MMD (grid)	GMSC	Latent	MLP	**0.03**

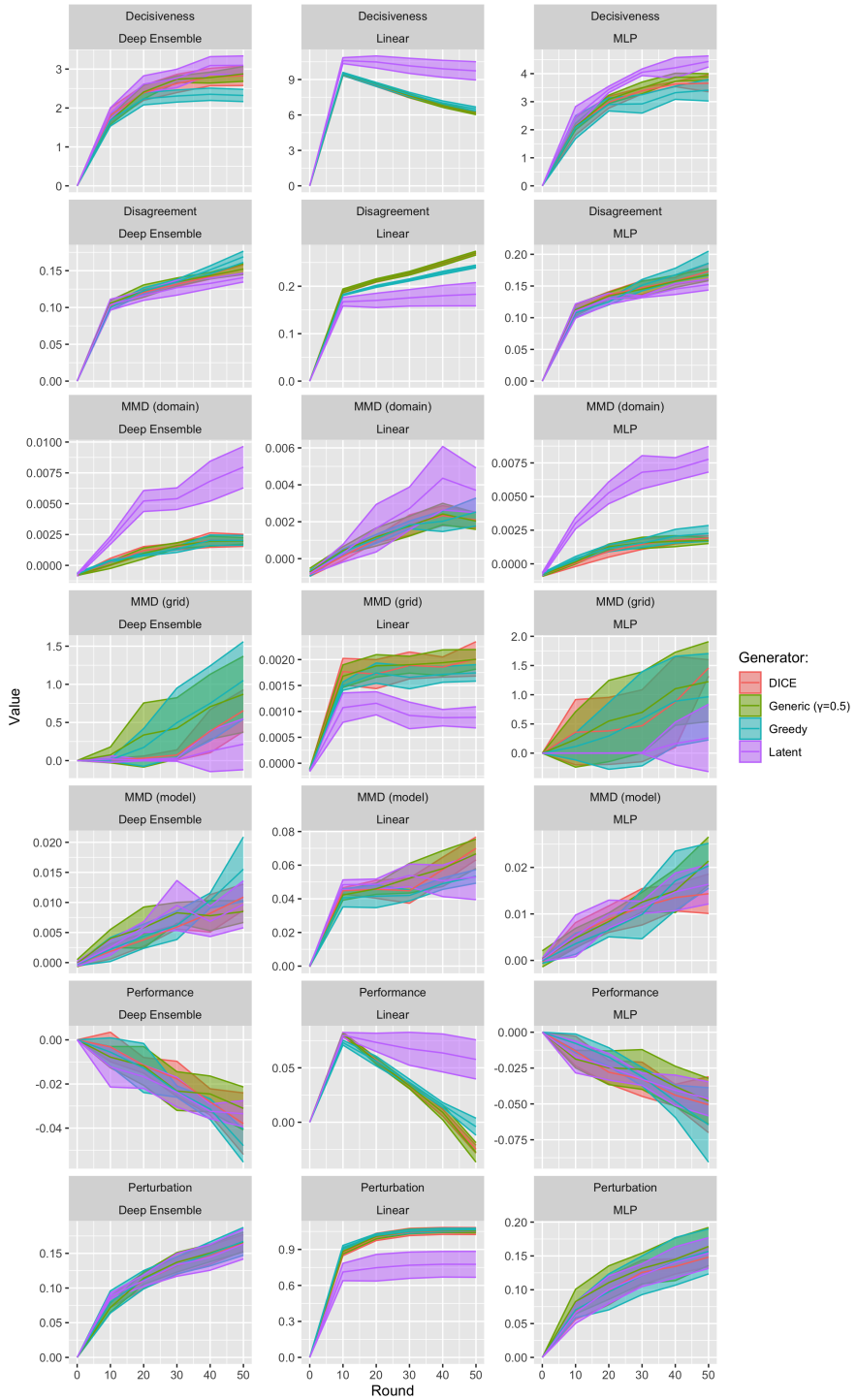


Figure D.9. Evolution of evaluation metrics over the course of the experiment. Data: California Housing.

D

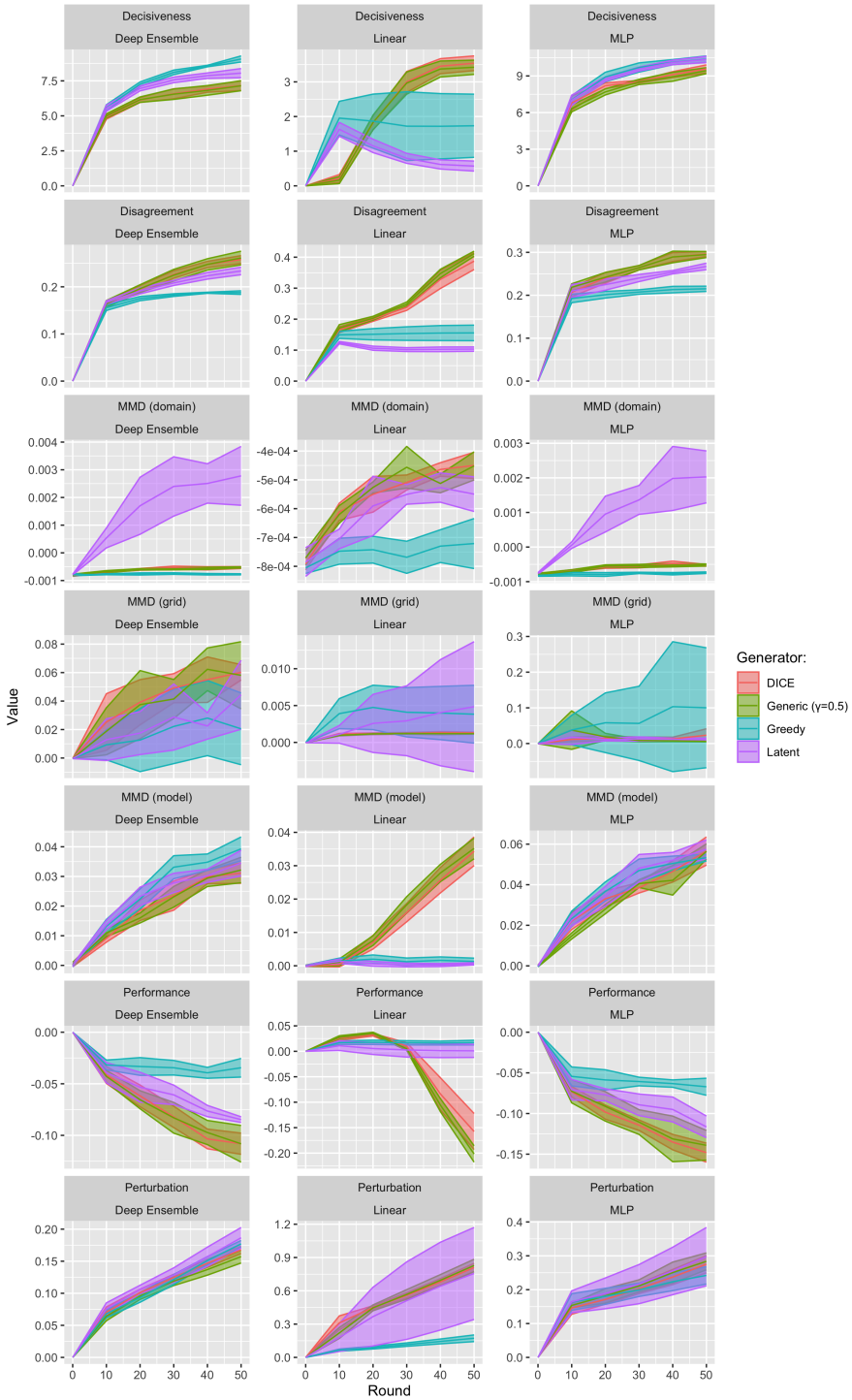


Figure D.10. Evolution of evaluation metrics over the course of the experiment. Data: Credit Default.

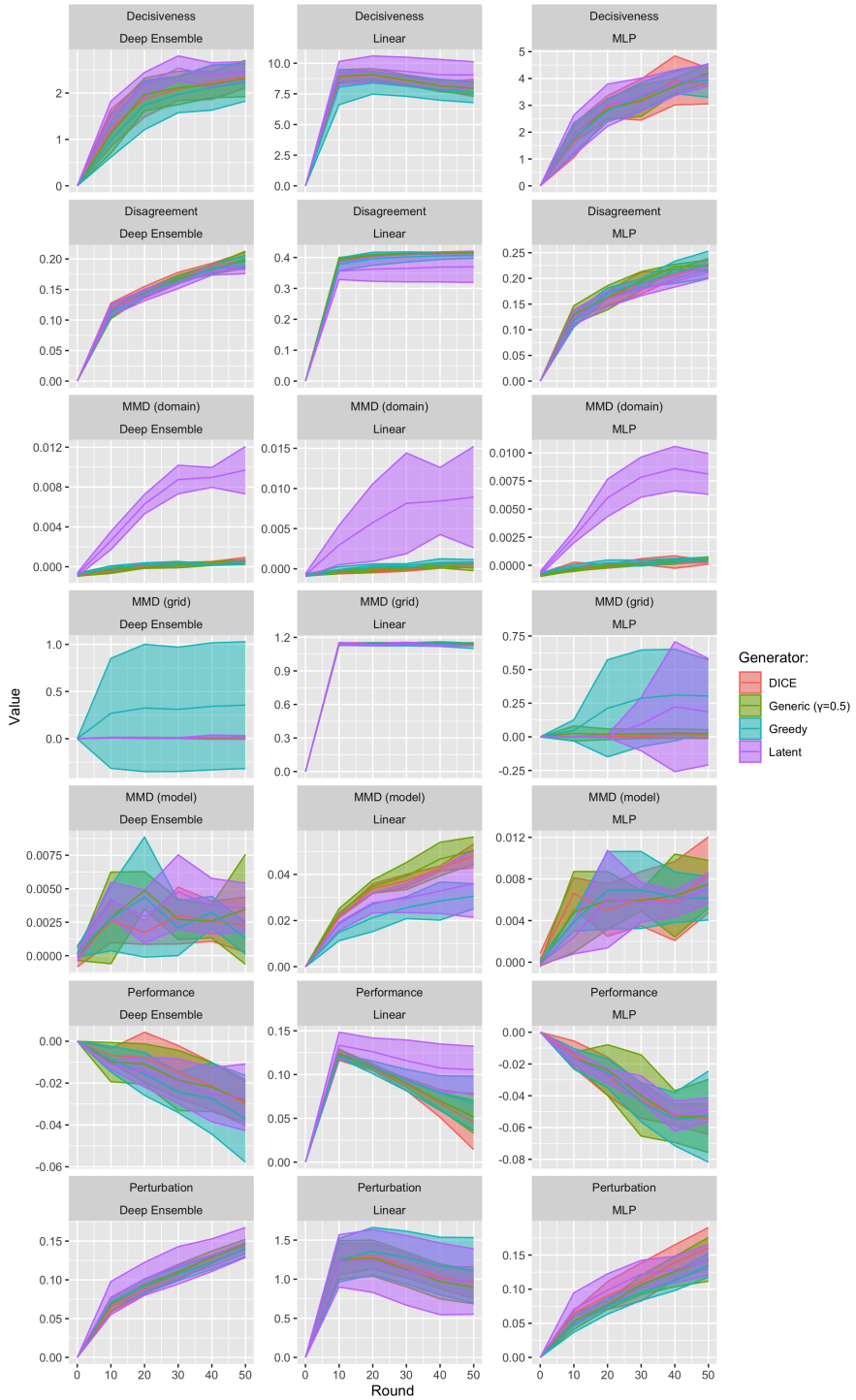


Figure D.11. Evolution of evaluation metrics over the course of the experiment. Data: GMSC.

D

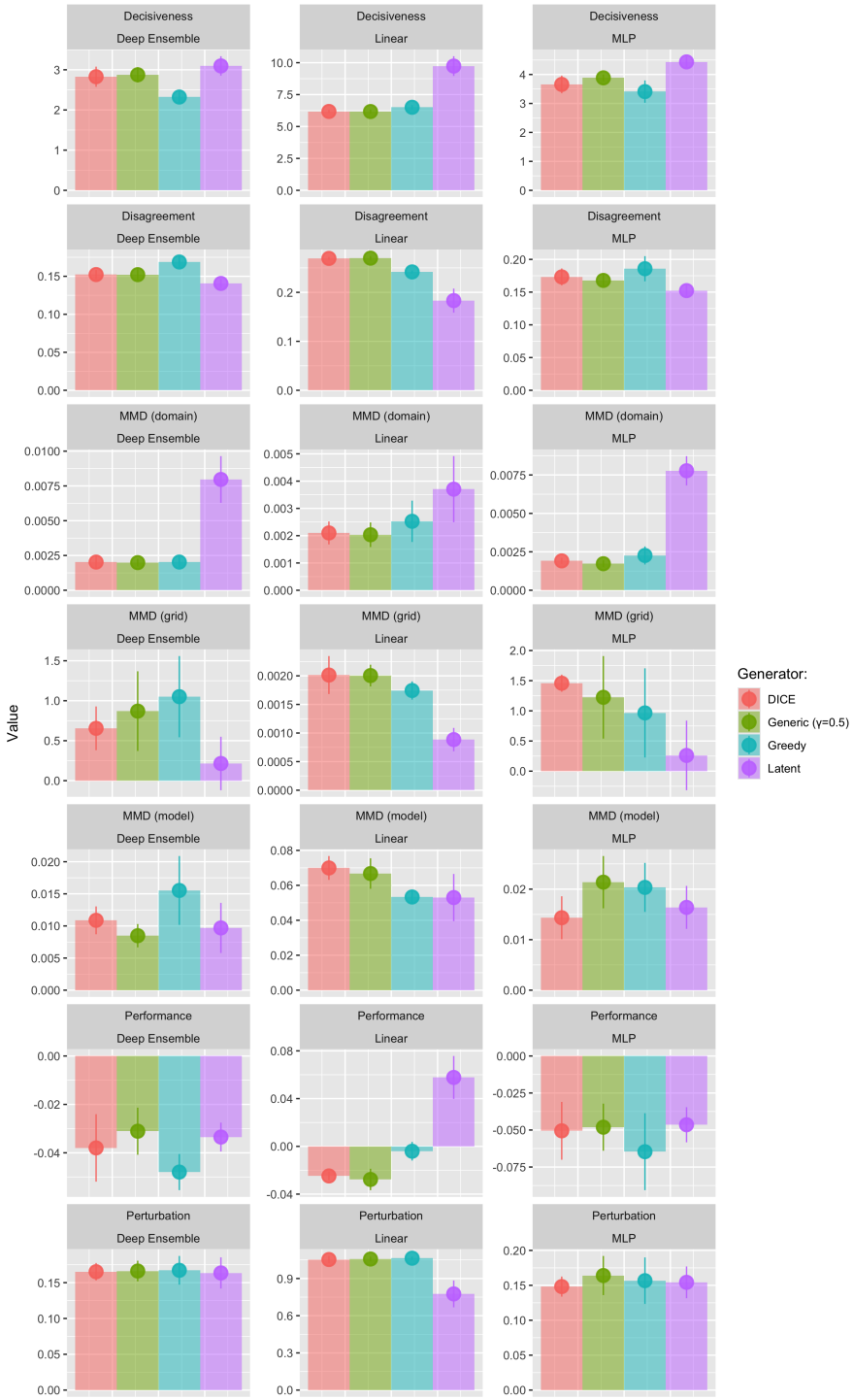


Figure D.12. Evaluation metrics at the end of the experiment. Data: California Housing.

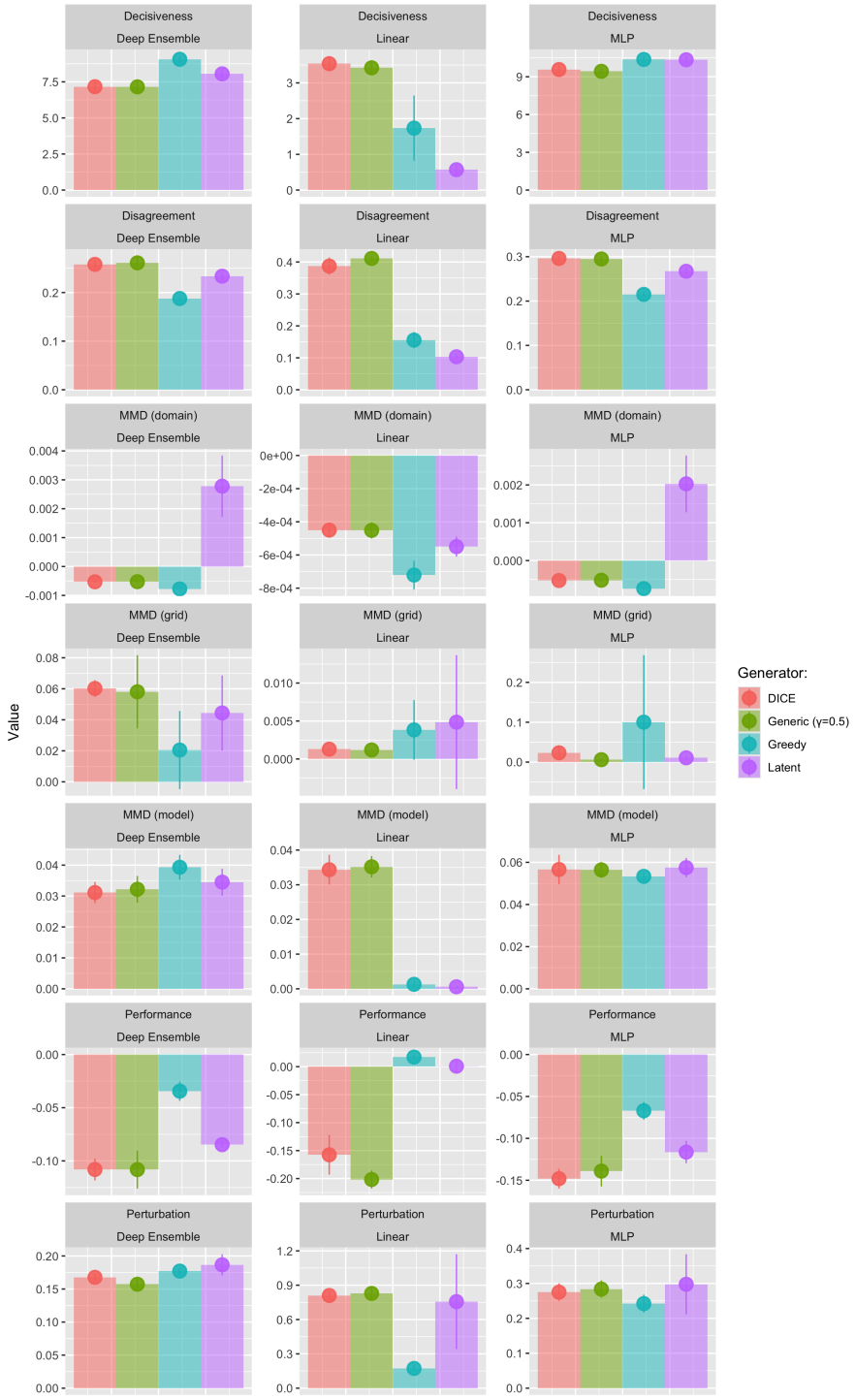


Figure D.13. Evaluation metrics at the end of the experiment. Data: Credit Default.

D

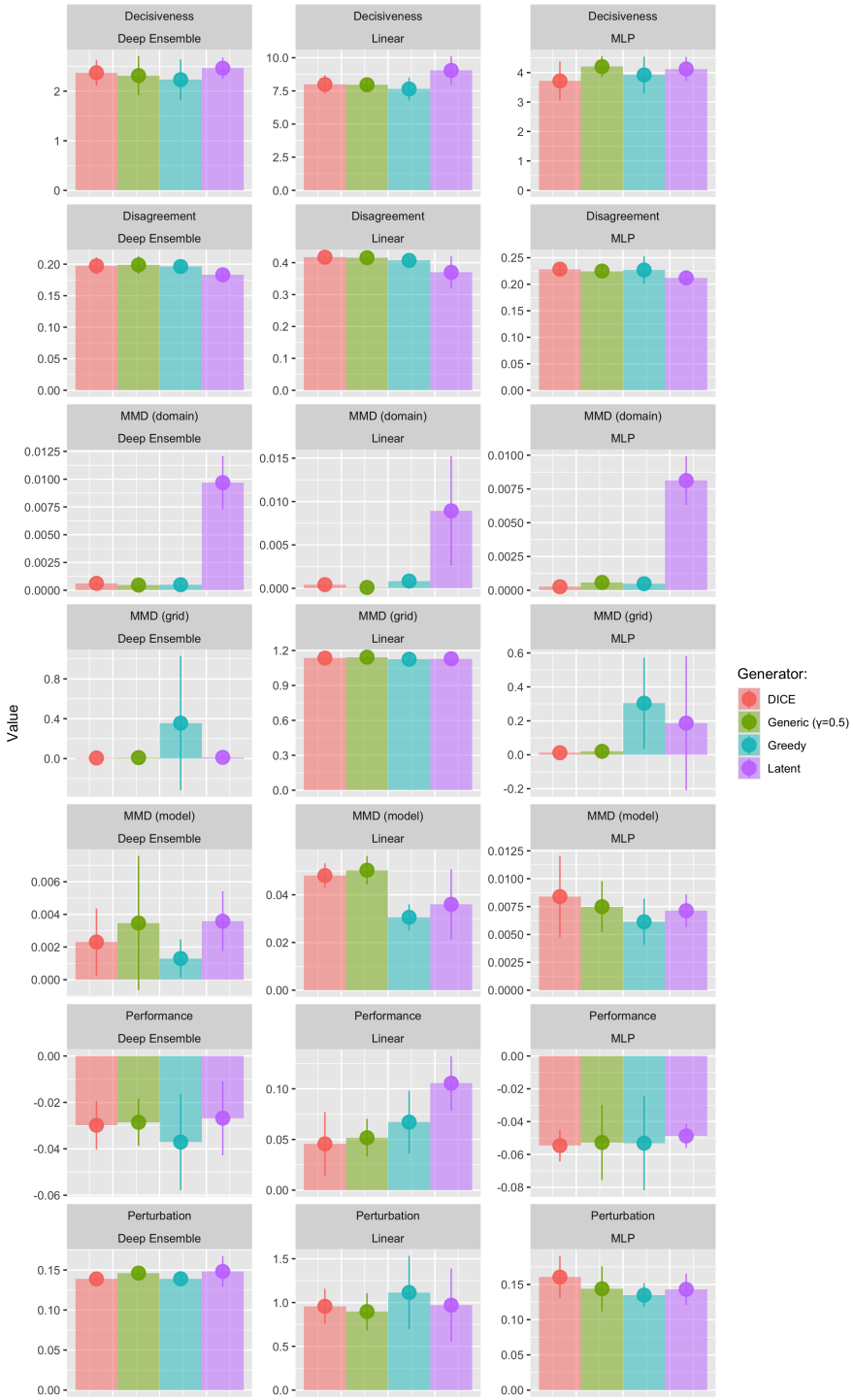


Figure D.14. Evaluation metrics at the end of the experiment. Data: GMSC.

D.3. DETAILED RESULTS: MITIGATION

D.3.1. LINE CHARTS

The evolution of the evaluation metrics over the course of the experiment is shown for different datasets in Figure D.15 to Figure D.21.

D.3.2. ERROR BAR CHARTS

The evaluation metrics at the end of the experiment are shown for different datasets in Figure D.22 to Figure D.28.

D.3.3. STATISTICAL SIGNIFICANCE

Table D.3 presents the tests for statistical significance of the estimated MMD metrics.

Table D.3. Tests for statistical significance of the estimated MMD metrics using mitigation strategies. We have highlighted p-values smaller than the significance level $\alpha = 0.05$ in bold. Data: Synthetic.

Metric	Data	Generator	Model	p-value
MMD	Circles	ClapROAR	Deep Ensemble	0.984
MMD	Circles	ClapROAR	Linear	1.0
MMD	Circles	ClapROAR	MLP	0.992
MMD	Circles	Generic (=0.5)	Deep Ensemble	0.99
MMD	Circles	Generic (=0.5)	Linear	1.0
MMD	Circles	Generic (=0.5)	MLP	0.994
MMD	Circles	Generic (=0.9)	Deep Ensemble	0.996
MMD	Circles	Generic (=0.9)	Linear	1.0
MMD	Circles	Generic (=0.9)	MLP	0.992
MMD	Circles	Gravitational	Deep Ensemble	0.998
MMD	Circles	Gravitational	Linear	1.0
MMD	Circles	Gravitational	MLP	0.998
MMD	Circles	Latent	Deep Ensemble	1.0
MMD	Circles	Latent	Linear	1.0
MMD	Circles	Latent	MLP	1.0
MMD	Linearly Separable	ClapROAR	Deep Ensemble	**0.0**
MMD	Linearly Separable	ClapROAR	Linear	**0.0**
MMD	Linearly Separable	ClapROAR	MLP	**0.0**
MMD	Linearly Separable	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Linearly Separable	Generic (=0.5)	Linear	**0.0**
MMD	Linearly Separable	Generic (=0.5)	MLP	**0.0**
MMD	Linearly Separable	Generic (=0.9)	Deep Ensemble	**0.0**

Continued below.

D

Metric	Data	Generator	Model	p-value
MMD	Linearly Separable	Generic (=0.9)	Linear	**0.0**
MMD	Linearly Separable	Generic (=0.9)	MLP	**0.0**
MMD	Linearly Separable	Gravitational	Deep Ensemble	0.05
MMD	Linearly Separable	Gravitational	Linear	0.092
MMD	Linearly Separable	Gravitational	MLP	0.078
MMD	Linearly Separable	Latent	Deep Ensemble	0.724
MMD	Linearly Separable	Latent	Linear	0.75
MMD	Linearly Separable	Latent	MLP	0.742
MMD	Moons	ClapROAR	Deep Ensemble	**0.0**
MMD	Moons	ClapROAR	Linear	**0.0**
MMD	Moons	ClapROAR	MLP	**0.0**
MMD	Moons	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Moons	Generic (=0.5)	Linear	**0.0**
MMD	Moons	Generic (=0.5)	MLP	**0.0**
MMD	Moons	Generic (=0.9)	Deep Ensemble	**0.0**
MMD	Moons	Generic (=0.9)	Linear	**0.0**
MMD	Moons	Generic (=0.9)	MLP	**0.0**
MMD	Moons	Gravitational	Deep Ensemble	**0.0**
MMD	Moons	Gravitational	Linear	**0.0**
MMD	Moons	Gravitational	MLP	**0.0**
MMD	Moons	Latent	Deep Ensemble	**0.0**
MMD	Moons	Latent	Linear	**0.0**
MMD	Moons	Latent	MLP	**0.0**
MMD	Overlapping	ClapROAR	Deep Ensemble	**0.0**
MMD	Overlapping	ClapROAR	Linear	**0.0**
MMD	Overlapping	ClapROAR	MLP	**0.0**
MMD	Overlapping	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Overlapping	Generic (=0.5)	Linear	**0.0**
MMD	Overlapping	Generic (=0.5)	MLP	**0.0**
MMD	Overlapping	Generic (=0.9)	Deep Ensemble	**0.0**
MMD	Overlapping	Generic (=0.9)	Linear	**0.0**
MMD	Overlapping	Generic (=0.9)	MLP	**0.0**
MMD	Overlapping	Gravitational	Deep Ensemble	**0.0**
MMD	Overlapping	Gravitational	Linear	**0.0**
MMD	Overlapping	Gravitational	MLP	**0.0**
MMD	Overlapping	Latent	Deep Ensemble	**0.0**
MMD	Overlapping	Latent	Linear	**0.0**
MMD	Overlapping	Latent	MLP	**0.0**
PP MMD	Circles	ClapROAR	Deep Ensemble	0.998
PP MMD	Circles	ClapROAR	Linear	0.996
PP MMD	Circles	ClapROAR	MLP	0.998
PP MMD	Circles	Generic (=0.5)	Deep Ensemble	0.998
PP MMD	Circles	Generic (=0.5)	Linear	0.8

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD	Circles	Generic (=0.5)	MLP	1.0
PP MMD	Circles	Generic (=0.9)	Deep Ensemble	0.998
PP MMD	Circles	Generic (=0.9)	Linear	0.996
PP MMD	Circles	Generic (=0.9)	MLP	1.0
PP MMD	Circles	Gravitational	Deep Ensemble	0.978
PP MMD	Circles	Gravitational	Linear	**0.0**
PP MMD	Circles	Gravitational	MLP	0.986
PP MMD	Circles	Latent	Deep Ensemble	1.0
PP MMD	Circles	Latent	Linear	**0.0**
PP MMD	Circles	Latent	MLP	0.998
PP MMD	Linearly Separable	ClapROAR	Deep Ensemble	0.962
PP MMD	Linearly Separable	ClapROAR	Linear	0.916
PP MMD	Linearly Separable	ClapROAR	MLP	0.958
PP MMD	Linearly Separable	Generic (=0.5)	Deep Ensemble	0.922
PP MMD	Linearly Separable	Generic (=0.5)	Linear	**0.0**
PP MMD	Linearly Separable	Generic (=0.5)	MLP	0.916
PP MMD	Linearly Separable	Generic (=0.9)	Deep Ensemble	0.968
PP MMD	Linearly Separable	Generic (=0.9)	Linear	0.376
PP MMD	Linearly Separable	Generic (=0.9)	MLP	0.968
PP MMD	Linearly Separable	Gravitational	Deep Ensemble	0.976
PP MMD	Linearly Separable	Gravitational	Linear	0.904
PP MMD	Linearly Separable	Gravitational	MLP	0.982
PP MMD	Linearly Separable	Latent	Deep Ensemble	0.862
PP MMD	Linearly Separable	Latent	Linear	0.428
PP MMD	Linearly Separable	Latent	MLP	0.83
PP MMD	Moons	ClapROAR	Deep Ensemble	0.966
PP MMD	Moons	ClapROAR	Linear	0.462
PP MMD	Moons	ClapROAR	MLP	0.956
PP MMD	Moons	Generic (=0.5)	Deep Ensemble	0.822
PP MMD	Moons	Generic (=0.5)	Linear	**0.0**
PP MMD	Moons	Generic (=0.5)	MLP	0.812
PP MMD	Moons	Generic (=0.9)	Deep Ensemble	0.818
PP MMD	Moons	Generic (=0.9)	Linear	0.086
PP MMD	Moons	Generic (=0.9)	MLP	0.87
PP MMD	Moons	Gravitational	Deep Ensemble	0.9775
PP MMD	Moons	Gravitational	Linear	0.446
PP MMD	Moons	Gravitational	MLP	0.984
PP MMD	Moons	Latent	Deep Ensemble	0.922
PP MMD	Moons	Latent	Linear	**0.008**
PP MMD	Moons	Latent	MLP	0.94
PP MMD	Overlapping	ClapROAR	Deep Ensemble	0.46
PP MMD	Overlapping	ClapROAR	Linear	0.178
PP MMD	Overlapping	ClapROAR	MLP	0.486

Continued below.

D

D

Metric	Data	Generator	Model	p-value
PP MMD	Overlapping	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	Overlapping	Generic (=0.5)	Linear	**0.0**
PP MMD	Overlapping	Generic (=0.5)	MLP	**0.004**
PP MMD	Overlapping	Generic (=0.9)	Deep Ensemble	0.122
PP MMD	Overlapping	Generic (=0.9)	Linear	0.066
PP MMD	Overlapping	Generic (=0.9)	MLP	0.13
PP MMD	Overlapping	Gravitational	Deep Ensemble	0.514
PP MMD	Overlapping	Gravitational	Linear	0.156
PP MMD	Overlapping	Gravitational	MLP	0.564
PP MMD	Overlapping	Latent	Deep Ensemble	**0.048**
PP MMD	Overlapping	Latent	Linear	**0.006**
PP MMD	Overlapping	Latent	MLP	**0.046**
PP MMD (grid)	Circles	ClapROAR	Deep Ensemble	0.984
PP MMD (grid)	Circles	ClapROAR	Linear	0.996
PP MMD (grid)	Circles	ClapROAR	MLP	0.99
PP MMD (grid)	Circles	Generic (=0.5)	Deep Ensemble	0.886
PP MMD (grid)	Circles	Generic (=0.5)	Linear	0.814
PP MMD (grid)	Circles	Generic (=0.5)	MLP	0.814
PP MMD (grid)	Circles	Generic (=0.9)	Deep Ensemble	0.84
PP MMD (grid)	Circles	Generic (=0.9)	Linear	0.988
PP MMD (grid)	Circles	Generic (=0.9)	MLP	0.932
PP MMD (grid)	Circles	Gravitational	Deep Ensemble	0.55
PP MMD (grid)	Circles	Gravitational	Linear	**0.0**
PP MMD (grid)	Circles	Gravitational	MLP	0.406
PP MMD (grid)	Circles	Latent	Deep Ensemble	0.996
PP MMD (grid)	Circles	Latent	Linear	**0.0**
PP MMD (grid)	Circles	Latent	MLP	0.99
PP MMD (grid)	Linearly Separable	ClapROAR	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	ClapROAR	Linear	**0.006**
PP MMD (grid)	Linearly Separable	ClapROAR	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.9)	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Generic (=0.9)	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Gravitational	Deep Ensemble	0.408
PP MMD (grid)	Linearly Separable	Gravitational	Linear	0.342
PP MMD (grid)	Linearly Separable	Gravitational	MLP	0.668
PP MMD (grid)	Linearly Separable	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Latent	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Latent	MLP	**0.0**
PP MMD (grid)	Moons	ClapROAR	Deep Ensemble	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD (grid)	Moons	ClapROAR	Linear	0.458
PP MMD (grid)	Moons	ClapROAR	MLP	**0.004**
PP MMD (grid)	Moons	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Moons	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Moons	Generic (=0.5)	MLP	**0.016**
PP MMD (grid)	Moons	Generic (=0.9)	Deep Ensemble	**0.006**
PP MMD (grid)	Moons	Generic (=0.9)	Linear	0.09
PP MMD (grid)	Moons	Generic (=0.9)	MLP	**0.03**
PP MMD (grid)	Moons	Gravitational	Deep Ensemble	0.4
PP MMD (grid)	Moons	Gravitational	Linear	0.456
PP MMD (grid)	Moons	Gravitational	MLP	0.426
PP MMD (grid)	Moons	Latent	Deep Ensemble	0.344
PP MMD (grid)	Moons	Latent	Linear	**0.008**
PP MMD (grid)	Moons	Latent	MLP	0.114
PP MMD (grid)	Overlapping	ClapROAR	Deep Ensemble	0.4075
PP MMD (grid)	Overlapping	ClapROAR	Linear	0.256
PP MMD (grid)	Overlapping	ClapROAR	MLP	0.298
PP MMD (grid)	Overlapping	Generic (=0.5)	Deep Ensemble	**0.002**
PP MMD (grid)	Overlapping	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Overlapping	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Overlapping	Generic (=0.9)	Deep Ensemble	0.154
PP MMD (grid)	Overlapping	Generic (=0.9)	Linear	0.104
PP MMD (grid)	Overlapping	Generic (=0.9)	MLP	0.116
PP MMD (grid)	Overlapping	Gravitational	Deep Ensemble	0.356
PP MMD (grid)	Overlapping	Gravitational	Linear	0.27
PP MMD (grid)	Overlapping	Gravitational	MLP	0.344
PP MMD (grid)	Overlapping	Latent	Deep Ensemble	0.324
PP MMD (grid)	Overlapping	Latent	Linear	**0.01**
PP MMD (grid)	Overlapping	Latent	MLP	0.204

Table D.4 presents the tests for statistical significance of the estimated MMD metrics.

Table D.4. Tests for statistical significance of the estimated MMD metrics using mitigation strategies. We have highlighted p-values smaller than the significance level $\alpha = 0.05$ in bold. Data: Real-World.

Metric	Data	Generator	Model	p-value
MMD	Cal Housing	ClapROAR	Deep Ensemble	**0.0**
MMD	Cal Housing	ClapROAR	Linear	**0.0**
MMD	Cal Housing	ClapROAR	MLP	**0.0**
MMD	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
MMD	Cal Housing	Generic (=0.5)	Linear	**0.0**

Continued below.

D

Metric	Data	Generator	Model	p-value
MMD	Cal Housing	Generic (=0.5)	MLP	**0.0**
MMD	Cal Housing	Generic (=0.9)	Deep Ensemble	**0.0**
MMD	Cal Housing	Generic (=0.9)	Linear	**0.0**
MMD	Cal Housing	Generic (=0.9)	MLP	**0.0**
MMD	Cal Housing	Gravitational	Deep Ensemble	**0.0**
MMD	Cal Housing	Gravitational	Linear	**0.0**
MMD	Cal Housing	Gravitational	MLP	**0.0**
MMD	Cal Housing	Latent	Deep Ensemble	**0.0**
MMD	Cal Housing	Latent	Linear	**0.0**
MMD	Cal Housing	Latent	MLP	**0.0**
MMD	Credit Default	ClapROAR	Deep Ensemble	1.0
MMD	Credit Default	ClapROAR	Linear	1.0
MMD	Credit Default	ClapROAR	MLP	1.0
MMD	Credit Default	Generic (=0.5)	Deep Ensemble	1.0
MMD	Credit Default	Generic (=0.5)	Linear	1.0
MMD	Credit Default	Generic (=0.5)	MLP	1.0
MMD	Credit Default	Generic (=0.9)	Deep Ensemble	1.0
MMD	Credit Default	Generic (=0.9)	Linear	1.0
MMD	Credit Default	Generic (=0.9)	MLP	1.0
MMD	Credit Default	Gravitational	Deep Ensemble	**0.0**
MMD	Credit Default	Gravitational	Linear	**0.0**
MMD	Credit Default	Gravitational	MLP	**0.0**
MMD	Credit Default	Latent	Deep Ensemble	**0.0**
MMD	Credit Default	Latent	Linear	0.8
MMD	Credit Default	Latent	MLP	**0.0**
MMD	GMSC	ClapROAR	Deep Ensemble	0.15
MMD	GMSC	ClapROAR	Linear	**0.0**
MMD	GMSC	ClapROAR	MLP	0.214
MMD	GMSC	Generic (=0.5)	Deep Ensemble	0.938
MMD	GMSC	Generic (=0.5)	Linear	0.856
MMD	GMSC	Generic (=0.5)	MLP	0.932
MMD	GMSC	Generic (=0.9)	Deep Ensemble	0.758
MMD	GMSC	Generic (=0.9)	Linear	**0.004**
MMD	GMSC	Generic (=0.9)	MLP	0.93
MMD	GMSC	Gravitational	Deep Ensemble	**0.0**
MMD	GMSC	Gravitational	Linear	**0.0**
MMD	GMSC	Gravitational	MLP	**0.0**
MMD	GMSC	Latent	Deep Ensemble	**0.0**
MMD	GMSC	Latent	Linear	**0.0**
MMD	GMSC	Latent	MLP	**0.0**
PP MMD	Cal Housing	ClapROAR	Deep Ensemble	**0.0**
PP MMD	Cal Housing	ClapROAR	Linear	**0.0**
PP MMD	Cal Housing	ClapROAR	MLP	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Generic (=0.5)	Linear	**0.0**
PP MMD	Cal Housing	Generic (=0.5)	MLP	**0.0**
PP MMD	Cal Housing	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Generic (=0.9)	Linear	**0.0**
PP MMD	Cal Housing	Generic (=0.9)	MLP	**0.0**
PP MMD	Cal Housing	Gravitational	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Gravitational	Linear	**0.0**
PP MMD	Cal Housing	Gravitational	MLP	**0.0**
PP MMD	Cal Housing	Latent	Deep Ensemble	**0.0**
PP MMD	Cal Housing	Latent	Linear	**0.0**
PP MMD	Cal Housing	Latent	MLP	**0.0**
PP MMD	Credit Default	ClapROAR	Deep Ensemble	**0.0**
PP MMD	Credit Default	ClapROAR	Linear	**0.0**
PP MMD	Credit Default	ClapROAR	MLP	**0.0**
PP MMD	Credit Default	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	Credit Default	Generic (=0.5)	Linear	**0.0**
PP MMD	Credit Default	Generic (=0.5)	MLP	**0.0**
PP MMD	Credit Default	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD	Credit Default	Generic (=0.9)	Linear	**0.0**
PP MMD	Credit Default	Generic (=0.9)	MLP	**0.0**
PP MMD	Credit Default	Gravitational	Deep Ensemble	**0.0**
PP MMD	Credit Default	Gravitational	Linear	**0.0**
PP MMD	Credit Default	Gravitational	MLP	**0.0**
PP MMD	Credit Default	Latent	Deep Ensemble	**0.0**
PP MMD	Credit Default	Latent	Linear	**0.0**
PP MMD	Credit Default	Latent	MLP	**0.0**
PP MMD	GMSC	ClapROAR	Deep Ensemble	**0.0**
PP MMD	GMSC	ClapROAR	Linear	**0.0**
PP MMD	GMSC	ClapROAR	MLP	**0.0**
PP MMD	GMSC	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD	GMSC	Generic (=0.5)	Linear	**0.0**
PP MMD	GMSC	Generic (=0.5)	MLP	**0.0**
PP MMD	GMSC	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD	GMSC	Generic (=0.9)	Linear	**0.0**
PP MMD	GMSC	Generic (=0.9)	MLP	**0.0**
PP MMD	GMSC	Gravitational	Deep Ensemble	**0.0**
PP MMD	GMSC	Gravitational	Linear	**0.0**
PP MMD	GMSC	Gravitational	MLP	**0.0**
PP MMD	GMSC	Latent	Deep Ensemble	**0.0**
PP MMD	GMSC	Latent	Linear	**0.0**
PP MMD	GMSC	Latent	MLP	**0.0**
PP MMD (grid)	Cal Housing	ClapROAR	Deep Ensemble	**0.044**

Continued below.

D

D

Metric	Data	Generator	Model	p-value
PP MMD (grid)	Cal Housing	ClapROAR	Linear	**0.004**
PP MMD (grid)	Cal Housing	ClapROAR	MLP	**0.012**
PP MMD (grid)	Cal Housing	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.9)	Deep Ensemble	**0.002**
PP MMD (grid)	Cal Housing	Generic (=0.9)	Linear	**0.0**
PP MMD (grid)	Cal Housing	Generic (=0.9)	MLP	**0.0**
PP MMD (grid)	Cal Housing	Gravitational	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	Gravitational	Linear	**0.014**
PP MMD (grid)	Cal Housing	Gravitational	MLP	0.0625
PP MMD (grid)	Cal Housing	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	Cal Housing	Latent	Linear	**0.002**
PP MMD (grid)	Cal Housing	Latent	MLP	**0.0**
PP MMD (grid)	Credit Default	ClapROAR	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	ClapROAR	Linear	**0.0**
PP MMD (grid)	Credit Default	ClapROAR	MLP	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.9)	Linear	**0.0**
PP MMD (grid)	Credit Default	Generic (=0.9)	MLP	**0.0**
PP MMD (grid)	Credit Default	Gravitational	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Gravitational	Linear	**0.0**
PP MMD (grid)	Credit Default	Gravitational	MLP	**0.0**
PP MMD (grid)	Credit Default	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	Credit Default	Latent	Linear	0.078
PP MMD (grid)	Credit Default	Latent	MLP	**0.0**
PP MMD (grid)	GMSC	ClapROAR	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	ClapROAR	Linear	**0.0**
PP MMD (grid)	GMSC	ClapROAR	MLP	**0.0**
PP MMD (grid)	GMSC	Generic (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Generic (=0.5)	Linear	**0.0**
PP MMD (grid)	GMSC	Generic (=0.5)	MLP	**0.0**
PP MMD (grid)	GMSC	Generic (=0.9)	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Generic (=0.9)	Linear	**0.0**
PP MMD (grid)	GMSC	Generic (=0.9)	MLP	**0.0**
PP MMD (grid)	GMSC	Gravitational	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Gravitational	Linear	**0.0**
PP MMD (grid)	GMSC	Gravitational	MLP	**0.0**
PP MMD (grid)	GMSC	Latent	Deep Ensemble	**0.0**
PP MMD (grid)	GMSC	Latent	Linear	**0.0**

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD (grid)	GMSC	Latent	MLP	**0.0**

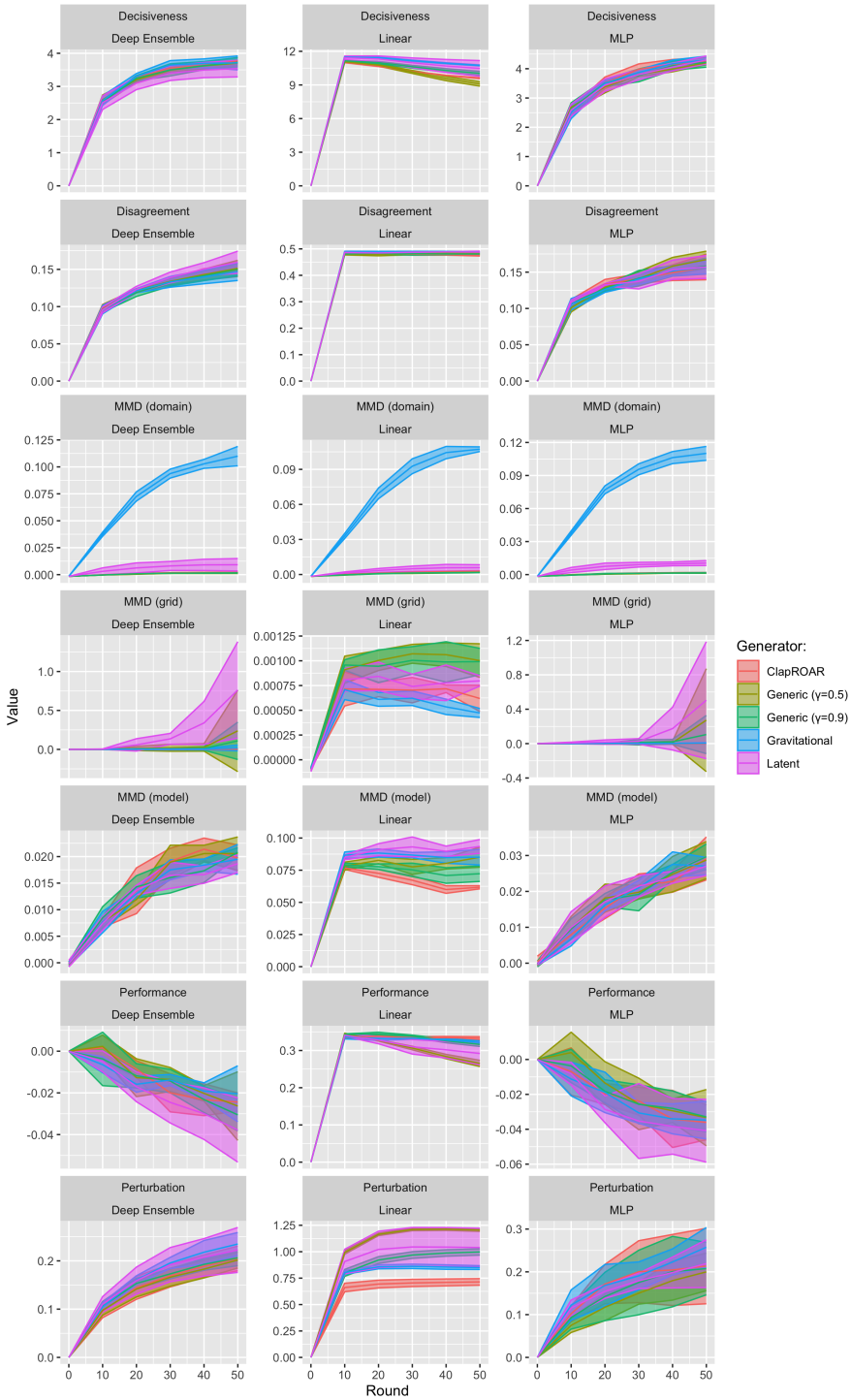


Figure D.15. Evolution of evaluation metrics over the course of the experiment. Data: California Housing.

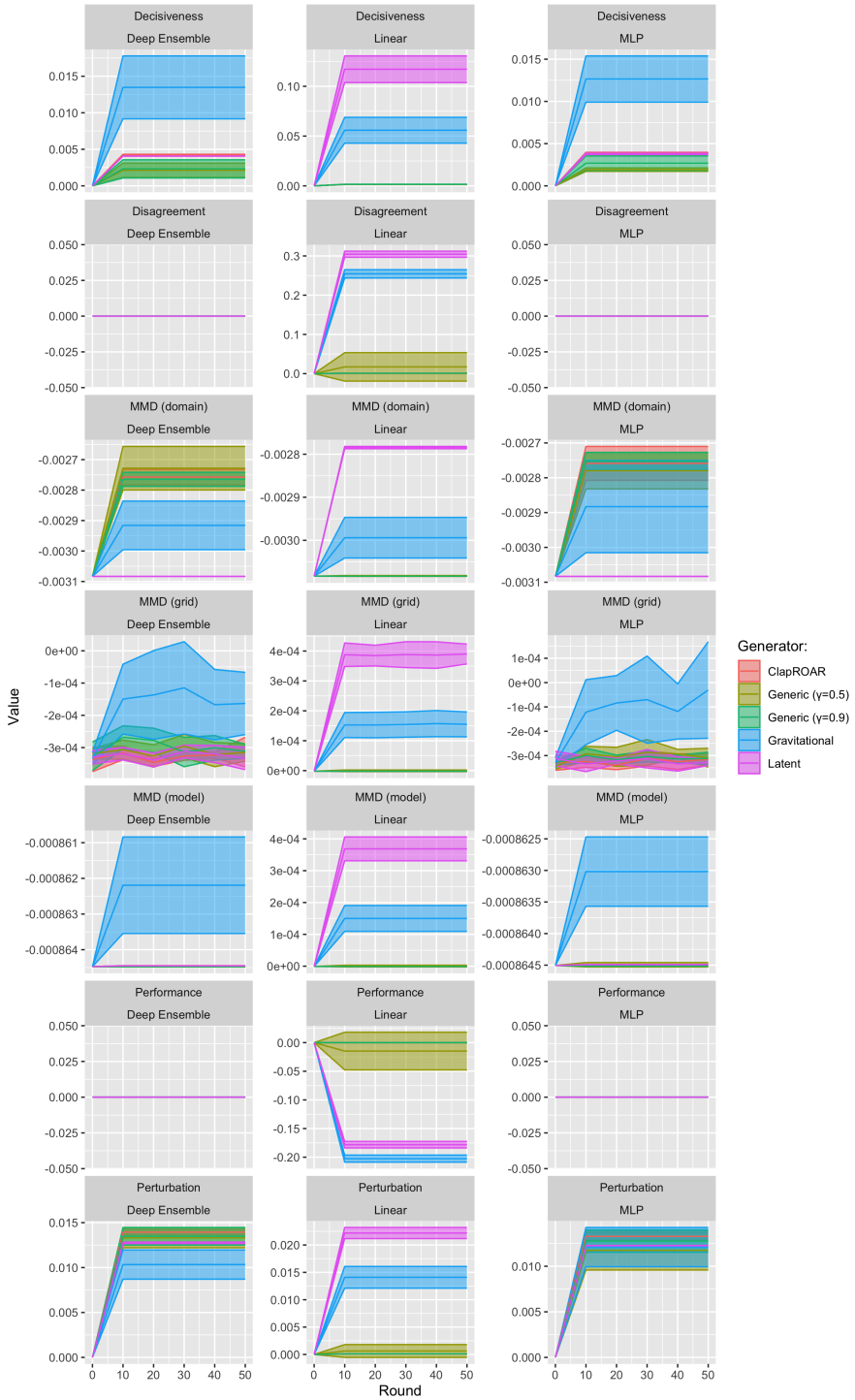


Figure D.16. Evolution of evaluation metrics over the course of the experiment. Data: Circles.

D

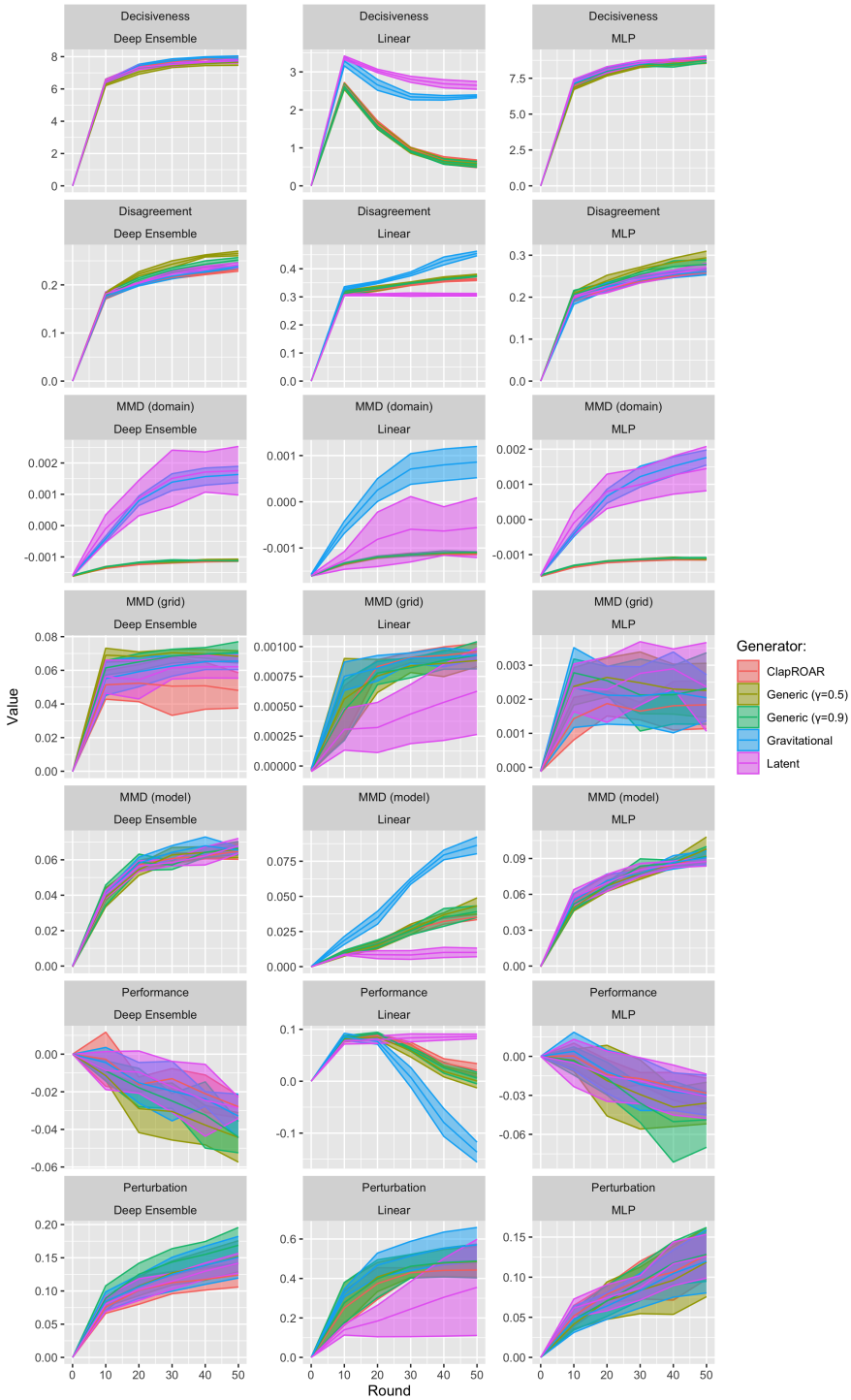


Figure D.17. Evolution of evaluation metrics over the course of the experiment. Data: Credit Default.

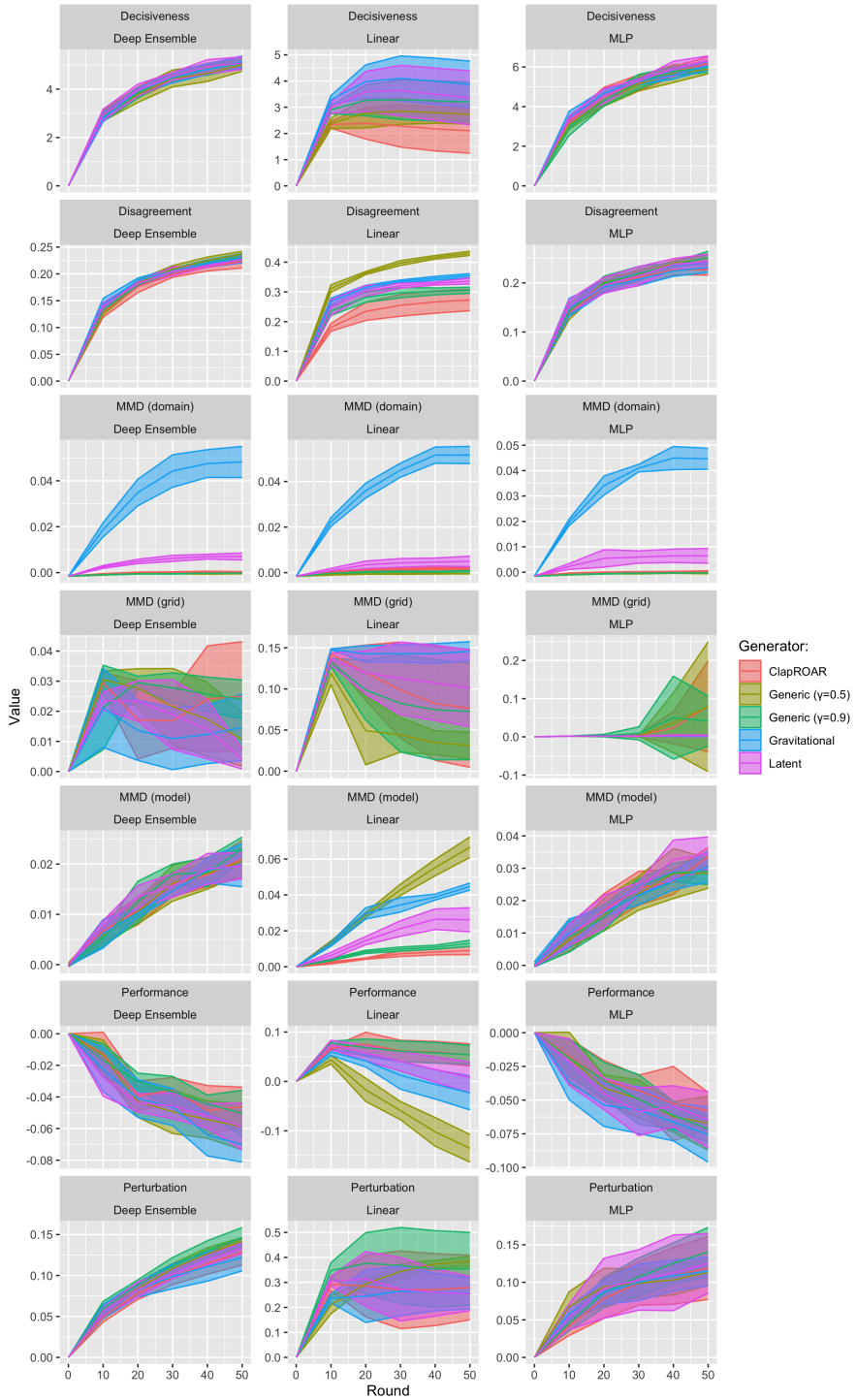


Figure D.18. Evolution of evaluation metrics over the course of the experiment. Data: GMSC.

D

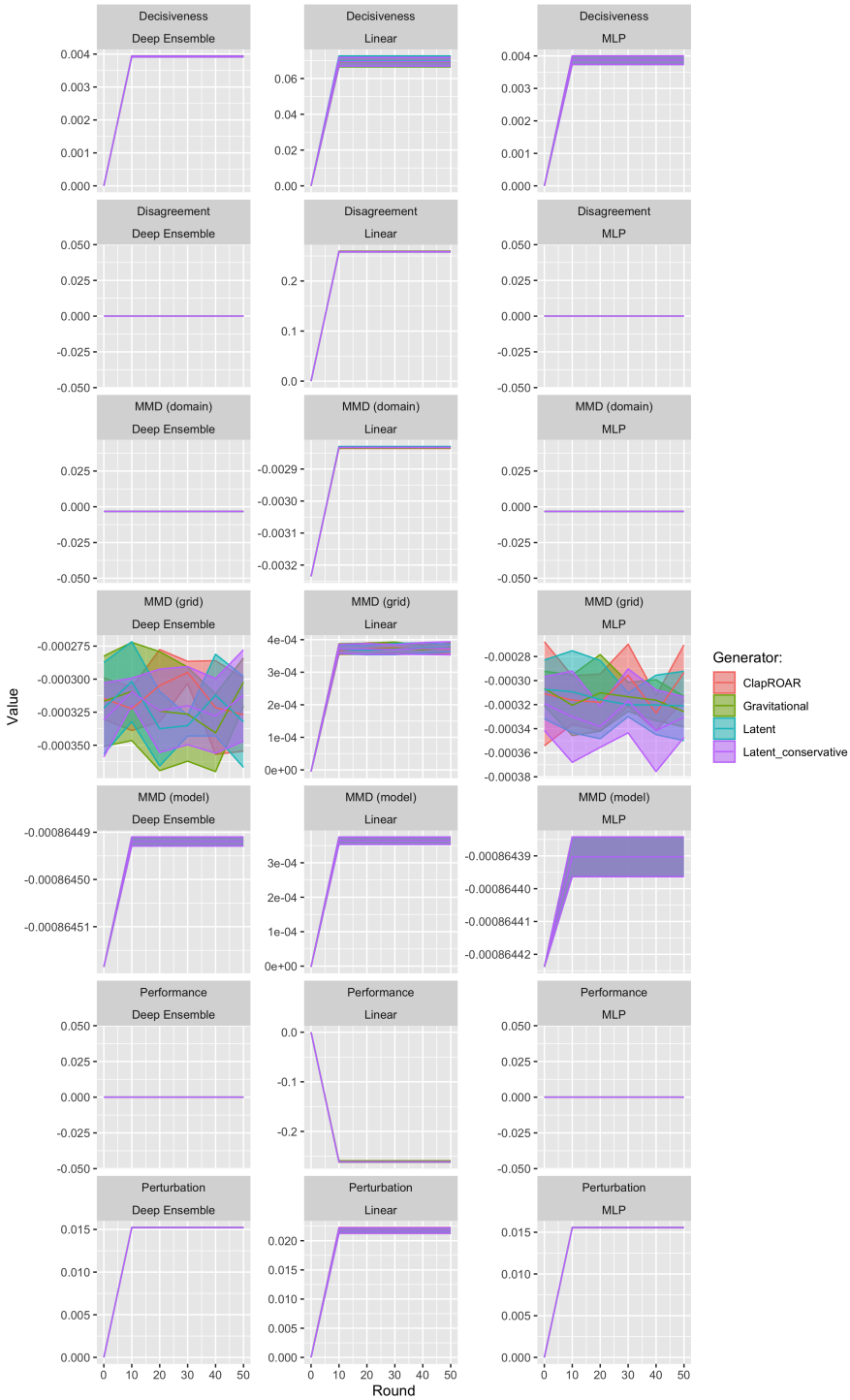


Figure D.19. Evolution of evaluation metrics over the course of the experiment. Data: Linearly Separable.

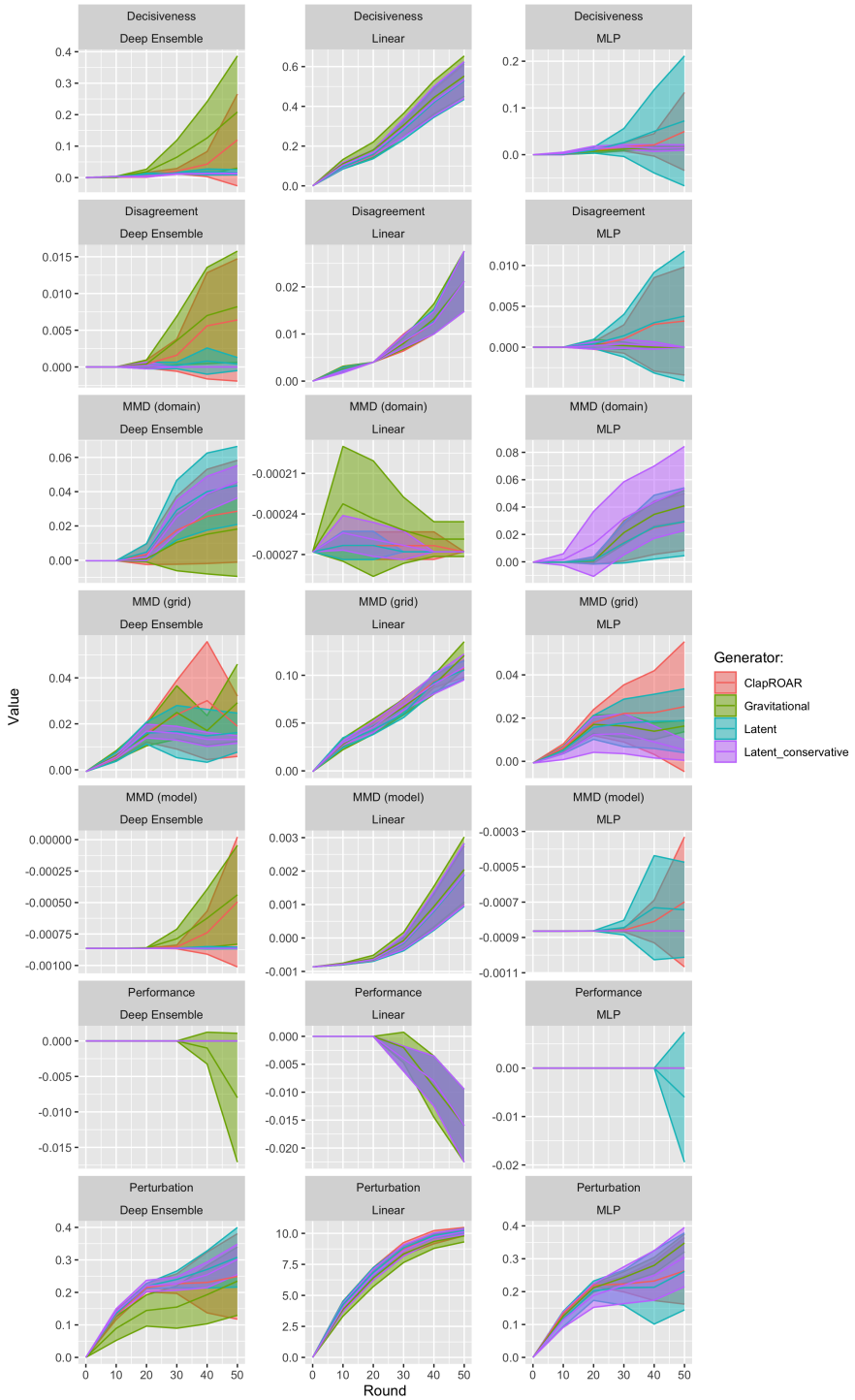


Figure D.20. Evolution of evaluation metrics over the course of the experiment. Data: Moons.

D

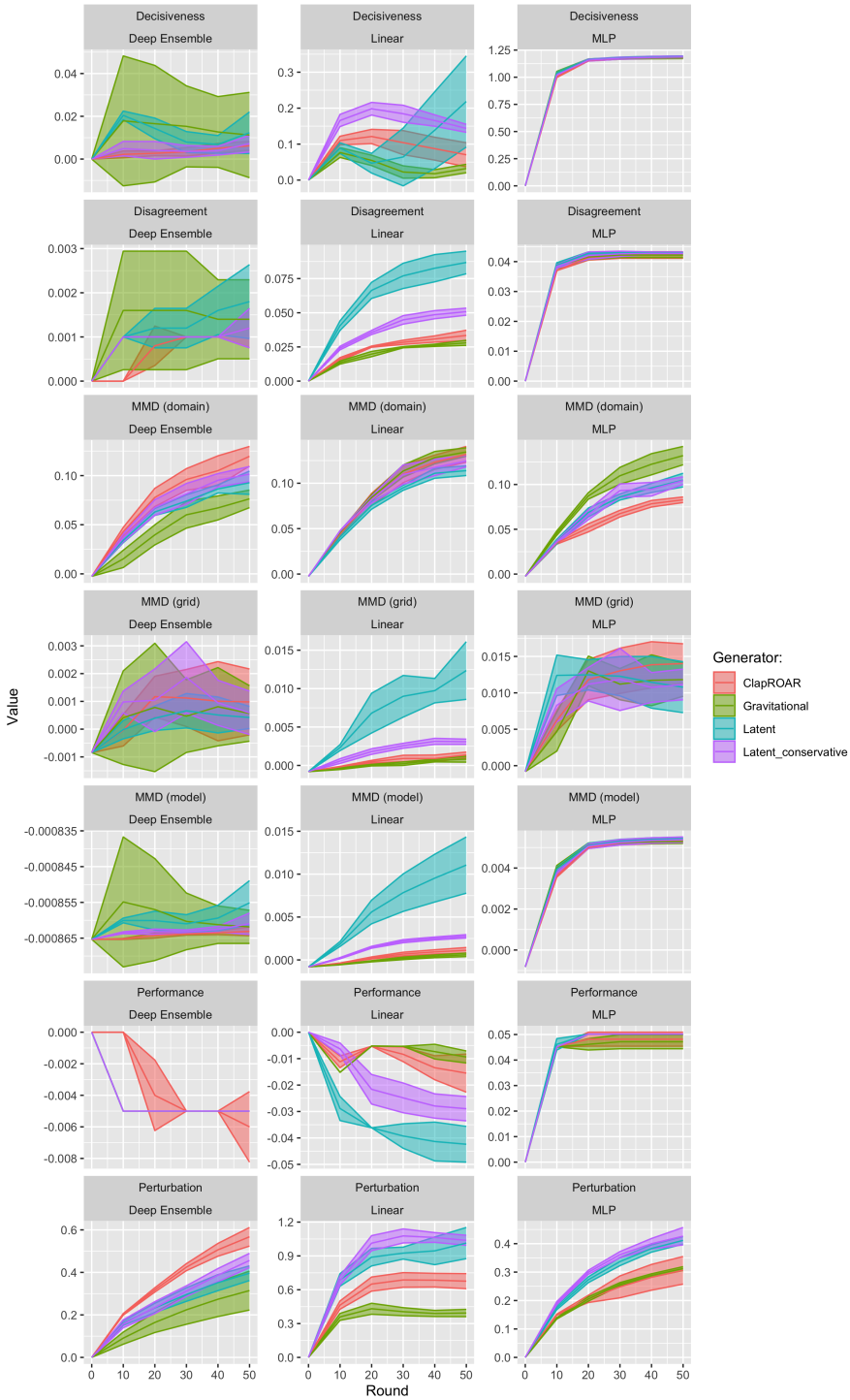


Figure D.21. Evolution of evaluation metrics over the course of the experiment. Data: Overlapping.

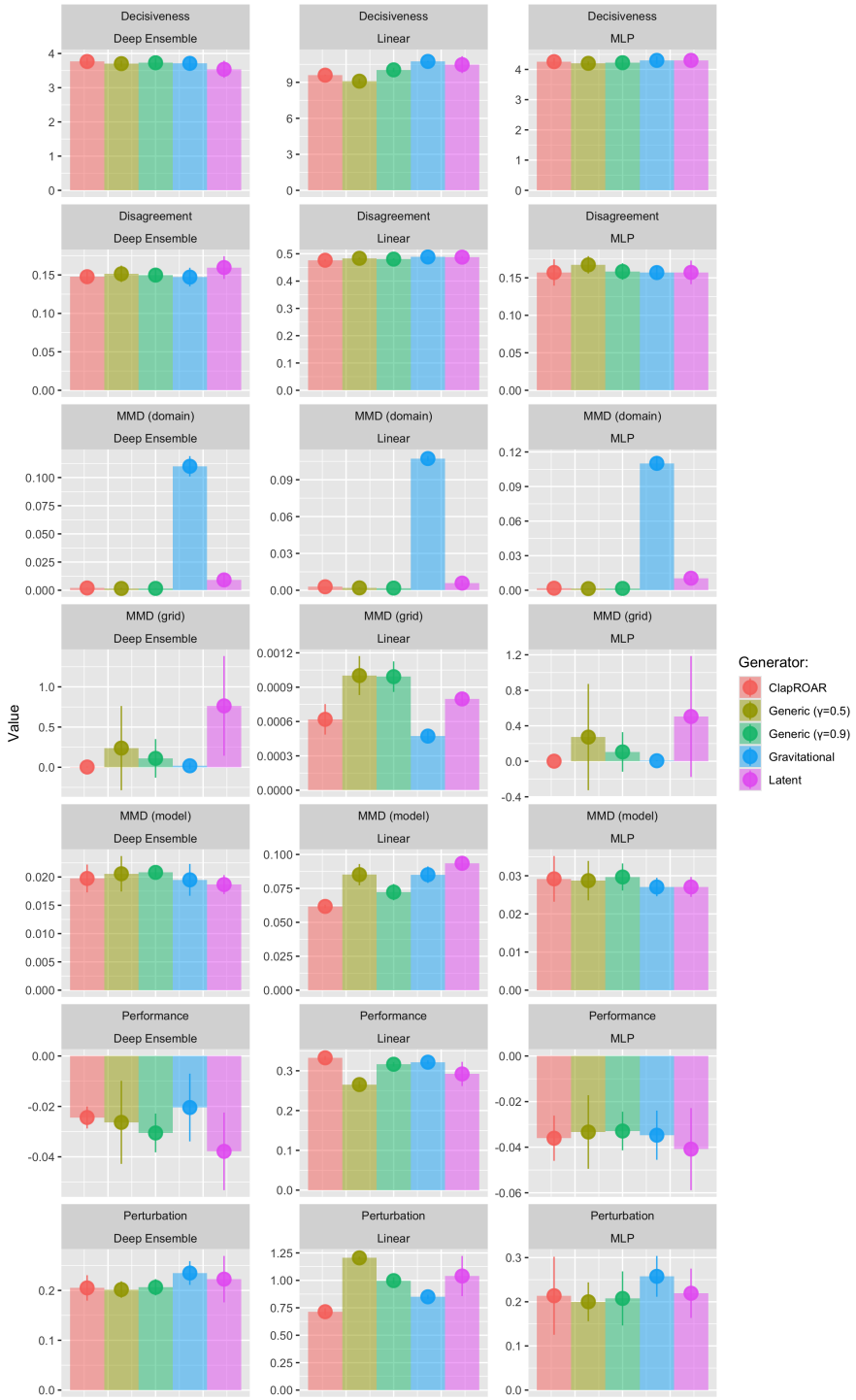


Figure D.22. Evaluation metrics at the end of the experiment. Data: California Housing.

D

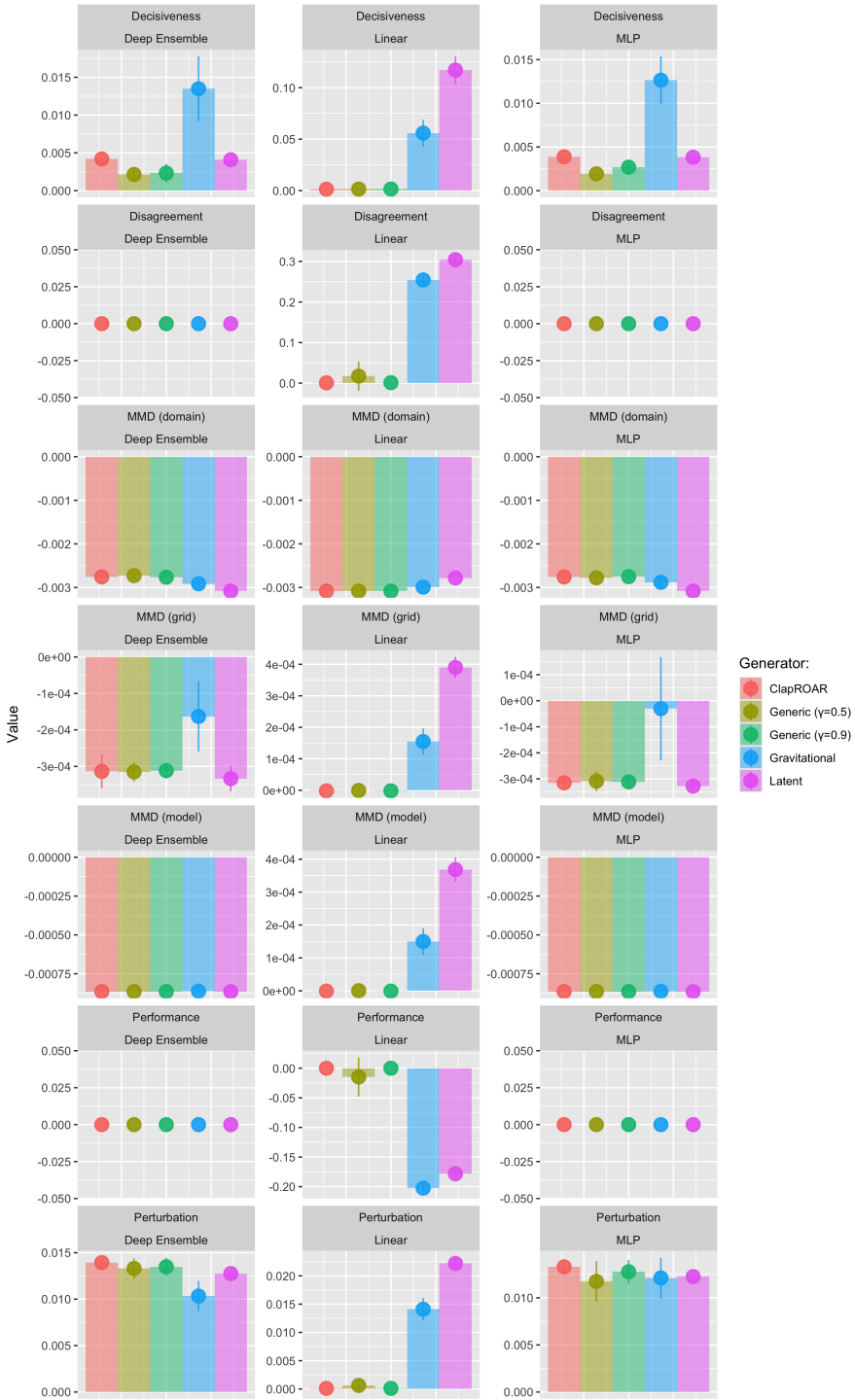


Figure D.23. Evaluation metrics at the end of the experiment. Data: Circles.

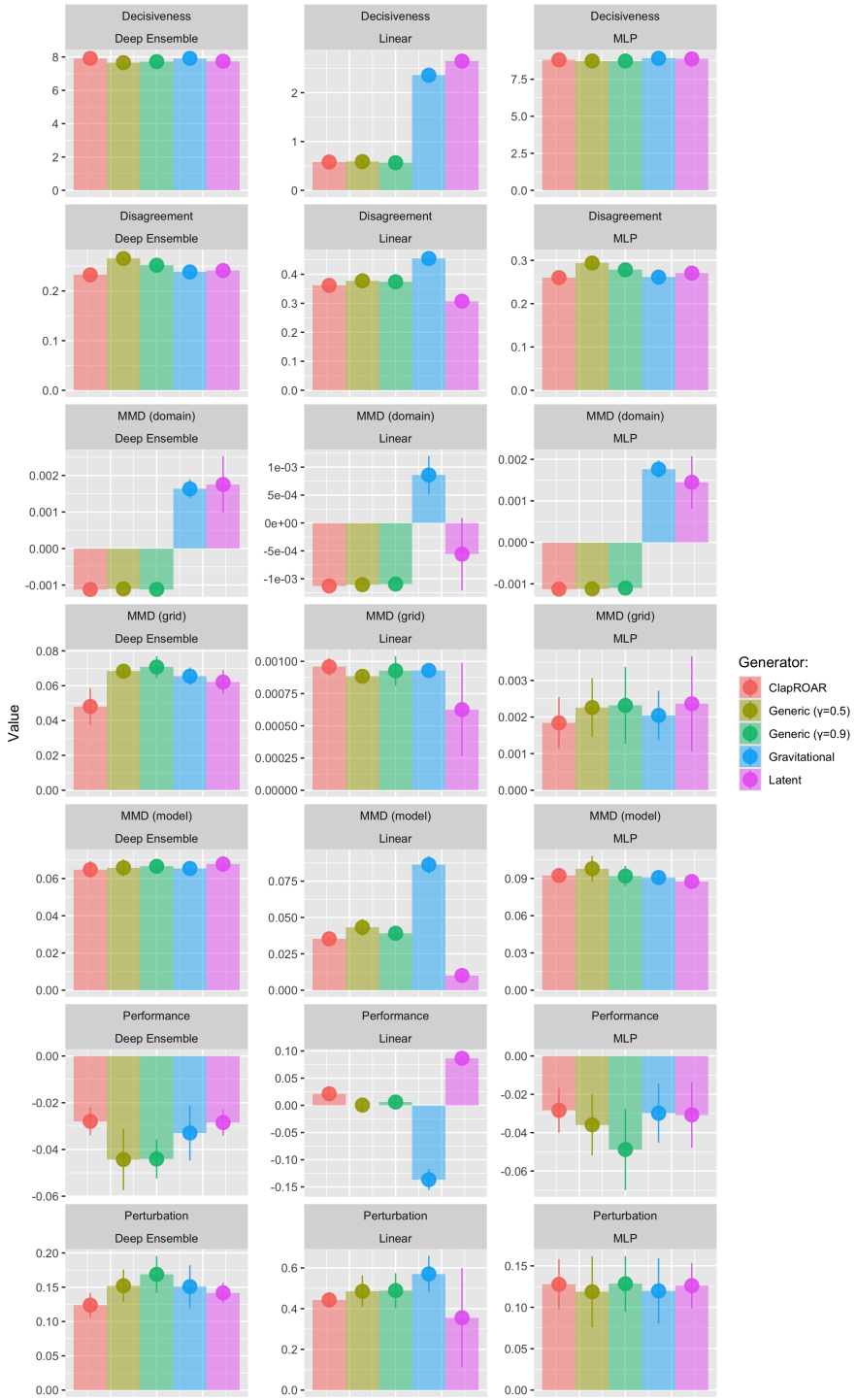


Figure D.24. Evaluation metrics at the end of the experiment. Data: Credit Default.

D

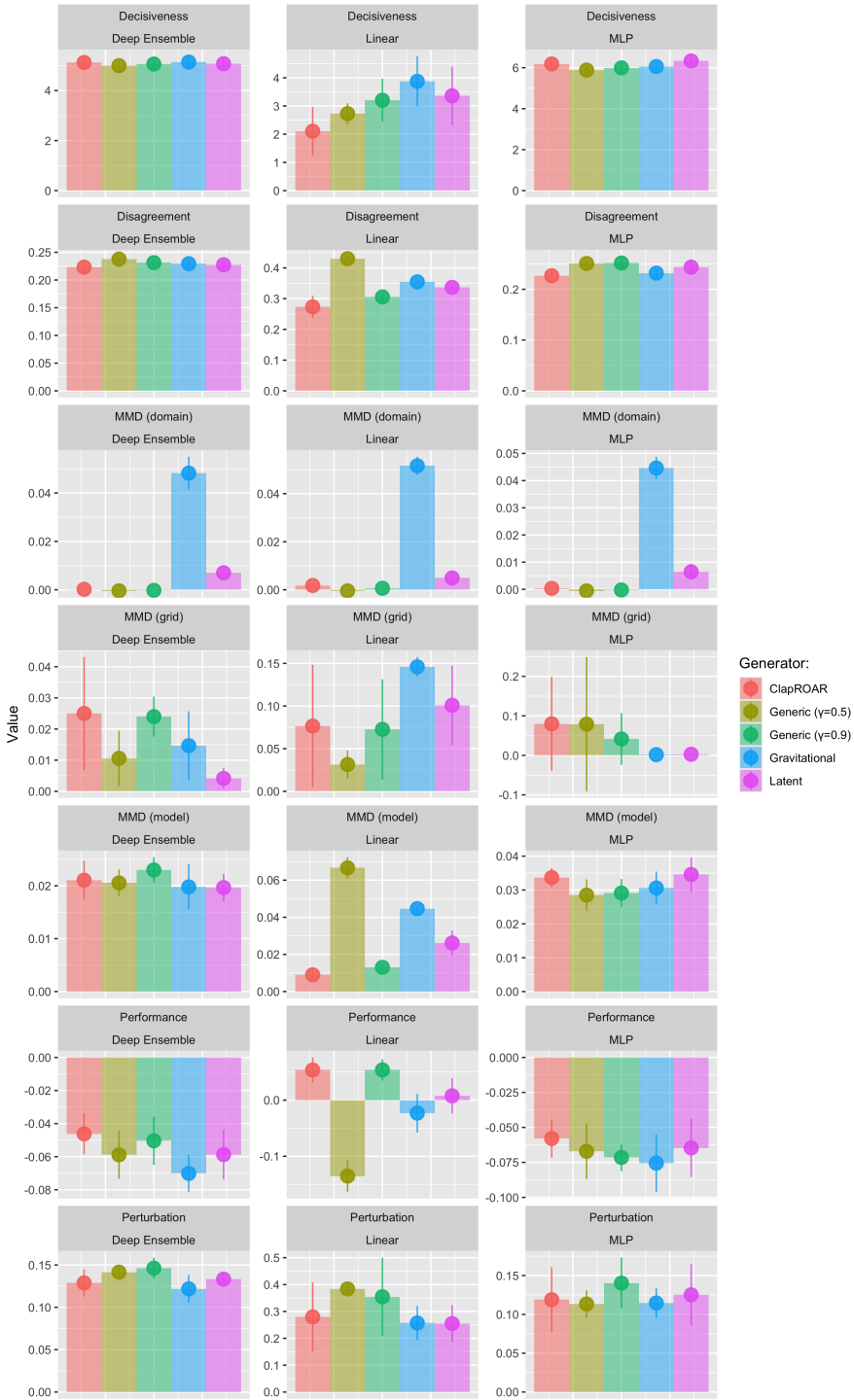


Figure D.25. Evaluation metrics at the end of the experiment. Data: GMSC.

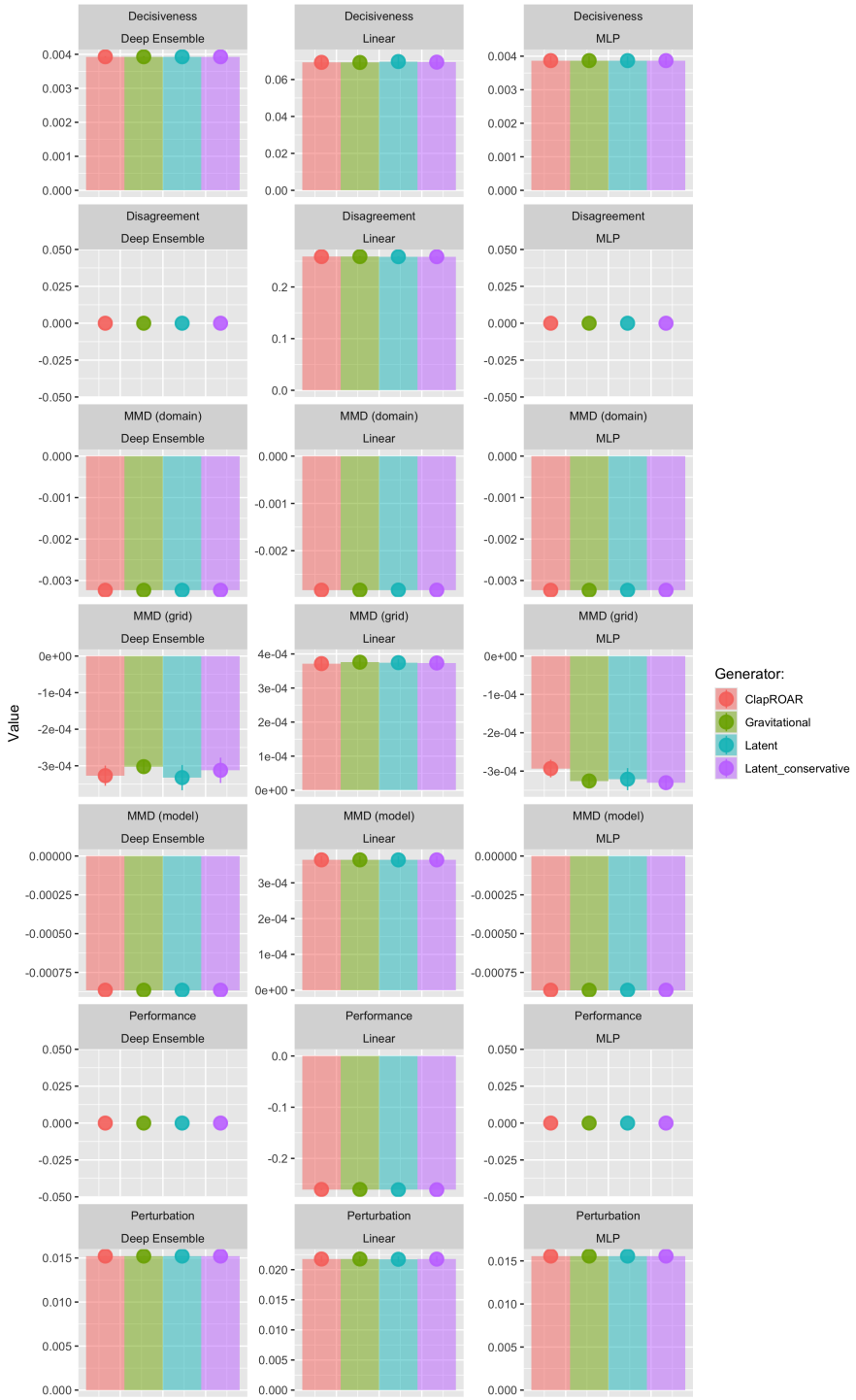


Figure D.26. Evaluation metrics at the end of the experiment. Data: Linearly Separable.

D

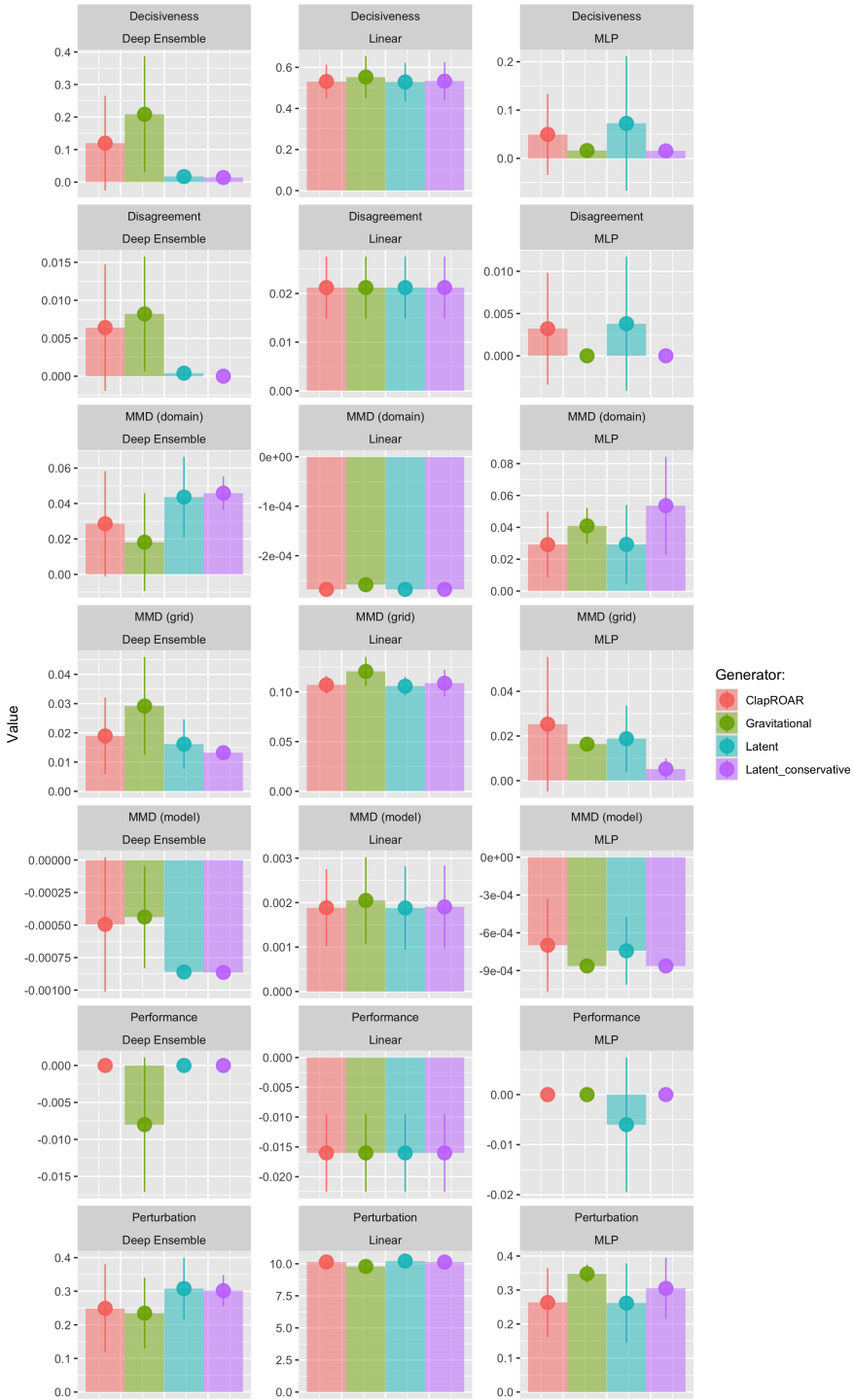


Figure D.27. Evaluation metrics at the end of the experiment. Data: Moons.

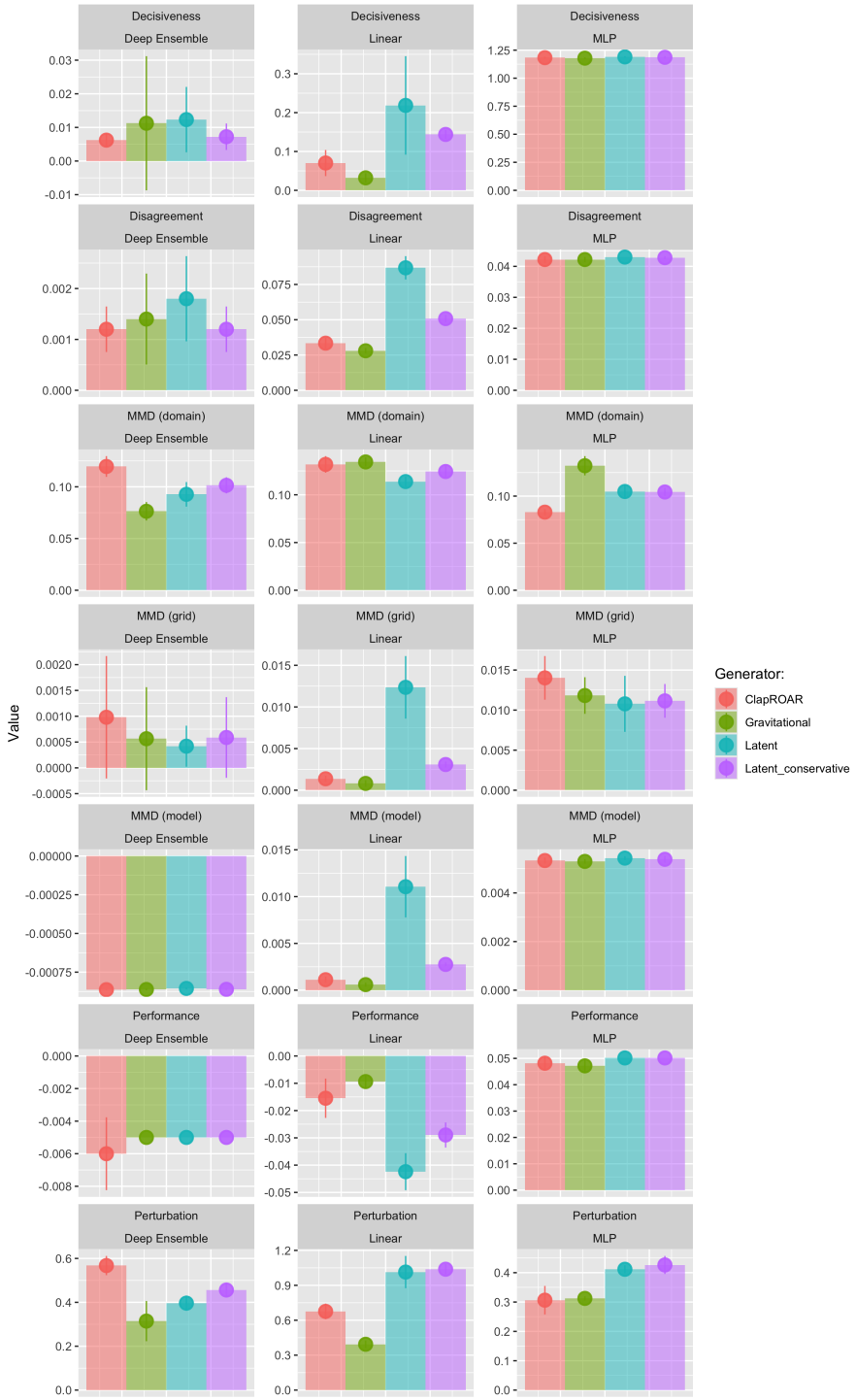


Figure D.28. Evaluation metrics at the end of the experiment. Data: Overlapping.

D

D.4. DETAILED RESULTS: MITIGATION WITH LATENT SPACE SEARCH

D.4.1. LINE CHARTS

The evolution of the evaluation metrics over the course of the experiment is shown for different datasets in Figure D.29 to Figure D.32.

D.4.2. ERROR BAR CHARTS

The evaluation metrics at the end of the experiment are shown for different datasets in Figure D.33 to Figure D.36.

D.4.3. STATISTICAL SIGNIFICANCE

Table D.5 presents the tests for statistical significance of the estimated MMD metrics.

Table D.5. Tests for statistical significance of the estimated MMD metrics using mitigation strategies with Latent Space Search. We have highlighted p-values smaller than the significance level $\alpha = 0.05$ in bold.

Metric	Data	Generator	Model	p-value
MMD	Circles	ClapROAR	Deep Ensemble	1.0
MMD	Circles	ClapROAR	Linear	0.994
MMD	Circles	ClapROAR	MLP	1.0
MMD	Circles	Gravitational	Deep Ensemble	0.998
MMD	Circles	Gravitational	Linear	1.0
MMD	Circles	Gravitational	MLP	1.0
MMD	Circles	Latent (=0.5)	Deep Ensemble	1.0
MMD	Circles	Latent (=0.5)	Linear	0.996
MMD	Circles	Latent (=0.5)	MLP	1.0
MMD	Circles	Latent (=0.9)	Deep Ensemble	1.0
MMD	Circles	Latent (=0.9)	Linear	0.996
MMD	Circles	Latent (=0.9)	MLP	1.0
MMD	Linearly Separable	ClapROAR	Deep Ensemble	0.334
MMD	Linearly Separable	ClapROAR	Linear	0.866
MMD	Linearly Separable	ClapROAR	MLP	0.168
MMD	Linearly Separable	Gravitational	Deep Ensemble	0.38
MMD	Linearly Separable	Gravitational	Linear	0.82
MMD	Linearly Separable	Gravitational	MLP	**0.0**
MMD	Linearly Separable	Latent (=0.5)	Deep Ensemble	**0.0**
MMD	Linearly Separable	Latent (=0.5)	Linear	0.892

Continued below.

Metric	Data	Generator	Model	p-value
MMD	Linearly Separable	Latent (=0.5)	MLP	0.126
MMD	Linearly Separable	Latent (=0.9)	Deep Ensemble	**0.0**
MMD	Linearly Separable	Latent (=0.9)	Linear	0.896
MMD	Linearly Separable	Latent (=0.9)	MLP	**0.0**
MMD	Moons	ClapROAR	Deep Ensemble	**0.0**
MMD	Moons	ClapROAR	Linear	**0.0**
MMD	Moons	ClapROAR	MLP	**0.0**
MMD	Moons	Gravitational	Deep Ensemble	**0.0**
MMD	Moons	Gravitational	Linear	**0.0**
MMD	Moons	Gravitational	MLP	**0.0**
MMD	Moons	Latent (=0.5)	Deep Ensemble	**0.0**
MMD	Moons	Latent (=0.5)	Linear	**0.0**
MMD	Moons	Latent (=0.5)	MLP	**0.0**
MMD	Moons	Latent (=0.9)	Deep Ensemble	**0.0**
MMD	Moons	Latent (=0.9)	Linear	**0.0**
MMD	Moons	Latent (=0.9)	MLP	**0.0**
MMD	Overlapping	ClapROAR	Deep Ensemble	**0.0**
MMD	Overlapping	ClapROAR	Linear	**0.0**
MMD	Overlapping	ClapROAR	MLP	**0.0**
MMD	Overlapping	Gravitational	Deep Ensemble	**0.0**
MMD	Overlapping	Gravitational	Linear	**0.0**
MMD	Overlapping	Gravitational	MLP	**0.0**
MMD	Overlapping	Latent (=0.5)	Deep Ensemble	**0.0**
MMD	Overlapping	Latent (=0.5)	Linear	**0.0**
MMD	Overlapping	Latent (=0.5)	MLP	**0.0**
MMD	Overlapping	Latent (=0.9)	Deep Ensemble	**0.0**
MMD	Overlapping	Latent (=0.9)	Linear	**0.0**
MMD	Overlapping	Latent (=0.9)	MLP	**0.0**
PP MMD	Circles	ClapROAR	Deep Ensemble	0.998
PP MMD	Circles	ClapROAR	Linear	**0.0**
PP MMD	Circles	ClapROAR	MLP	1.0
PP MMD	Circles	Gravitational	Deep Ensemble	1.0
PP MMD	Circles	Gravitational	Linear	**0.0**
PP MMD	Circles	Gravitational	MLP	1.0
PP MMD	Circles	Latent (=0.5)	Deep Ensemble	0.998
PP MMD	Circles	Latent (=0.5)	Linear	**0.0**
PP MMD	Circles	Latent (=0.5)	MLP	0.9975
PP MMD	Circles	Latent (=0.9)	Deep Ensemble	0.998
PP MMD	Circles	Latent (=0.9)	Linear	**0.0**
PP MMD	Circles	Latent (=0.9)	MLP	0.998
PP MMD	Linearly Separable	ClapROAR	Deep Ensemble	0.698
PP MMD	Linearly Separable	ClapROAR	Linear	0.094
PP MMD	Linearly Separable	ClapROAR	MLP	0.826

Continued below.

D

D

Metric	Data	Generator	Model	p-value
PP MMD	Linearly Separable	Gravitational	Deep Ensemble	0.616
PP MMD	Linearly Separable	Gravitational	Linear	0.096
PP MMD	Linearly Separable	Gravitational	MLP	0.962
PP MMD	Linearly Separable	Latent (=0.5)	Deep Ensemble	0.948
PP MMD	Linearly Separable	Latent (=0.5)	Linear	0.094
PP MMD	Linearly Separable	Latent (=0.5)	MLP	0.85
PP MMD	Linearly Separable	Latent (=0.9)	Deep Ensemble	0.96
PP MMD	Linearly Separable	Latent (=0.9)	Linear	0.072
PP MMD	Linearly Separable	Latent (=0.9)	MLP	0.966
PP MMD	Moons	ClapROAR	Deep Ensemble	0.962
PP MMD	Moons	ClapROAR	Linear	0.134
PP MMD	Moons	ClapROAR	MLP	**0.005**
PP MMD	Moons	Gravitational	Deep Ensemble	0.966
PP MMD	Moons	Gravitational	Linear	0.2075
PP MMD	Moons	Gravitational	MLP	**0.01**
PP MMD	Moons	Latent (=0.5)	Deep Ensemble	0.9075
PP MMD	Moons	Latent (=0.5)	Linear	**0.0**
PP MMD	Moons	Latent (=0.5)	MLP	**0.006**
PP MMD	Moons	Latent (=0.9)	Deep Ensemble	0.93
PP MMD	Moons	Latent (=0.9)	Linear	**0.0275**
PP MMD	Moons	Latent (=0.9)	MLP	**0.002**
PP MMD	Overlapping	ClapROAR	Deep Ensemble	0.412
PP MMD	Overlapping	ClapROAR	Linear	0.13
PP MMD	Overlapping	ClapROAR	MLP	0.34
PP MMD	Overlapping	Gravitational	Deep Ensemble	0.544
PP MMD	Overlapping	Gravitational	Linear	0.238
PP MMD	Overlapping	Gravitational	MLP	0.662
PP MMD	Overlapping	Latent (=0.5)	Deep Ensemble	**0.046**
PP MMD	Overlapping	Latent (=0.5)	Linear	**0.0**
PP MMD	Overlapping	Latent (=0.5)	MLP	0.07
PP MMD	Overlapping	Latent (=0.9)	Deep Ensemble	0.196
PP MMD	Overlapping	Latent (=0.9)	Linear	**0.046**
PP MMD	Overlapping	Latent (=0.9)	MLP	0.132
PP MMD (grid)	Circles	ClapROAR	Deep Ensemble	0.994
PP MMD (grid)	Circles	ClapROAR	Linear	**0.0**
PP MMD (grid)	Circles	ClapROAR	MLP	0.996
PP MMD (grid)	Circles	Gravitational	Deep Ensemble	0.992
PP MMD (grid)	Circles	Gravitational	Linear	**0.0**
PP MMD (grid)	Circles	Gravitational	MLP	0.996
PP MMD (grid)	Circles	Latent (=0.5)	Deep Ensemble	0.998
PP MMD (grid)	Circles	Latent (=0.5)	Linear	**0.0**
PP MMD (grid)	Circles	Latent (=0.5)	MLP	0.9925
PP MMD (grid)	Circles	Latent (=0.9)	Deep Ensemble	0.994

Continued below.

Metric	Data	Generator	Model	p-value
PP MMD (grid)	Circles	Latent (=0.9)	Linear	**0.0**
PP MMD (grid)	Circles	Latent (=0.9)	MLP	0.988
PP MMD (grid)	Linearly Separable	ClapROAR	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	ClapROAR	Linear	**0.0**
PP MMD (grid)	Linearly Separable	ClapROAR	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Gravitational	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Gravitational	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Gravitational	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.5)	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.5)	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.5)	MLP	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.9)	Deep Ensemble	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.9)	Linear	**0.0**
PP MMD (grid)	Linearly Separable	Latent (=0.9)	MLP	**0.044**
PP MMD (grid)	Moons	ClapROAR	Deep Ensemble	0.128
PP MMD (grid)	Moons	ClapROAR	Linear	0.072
PP MMD (grid)	Moons	ClapROAR	MLP	**0.0**
PP MMD (grid)	Moons	Gravitational	Deep Ensemble	0.2
PP MMD (grid)	Moons	Gravitational	Linear	0.1525
PP MMD (grid)	Moons	Gravitational	MLP	**0.0**
PP MMD (grid)	Moons	Latent (=0.5)	Deep Ensemble	0.22
PP MMD (grid)	Moons	Latent (=0.5)	Linear	**0.0**
PP MMD (grid)	Moons	Latent (=0.5)	MLP	**0.0**
PP MMD (grid)	Moons	Latent (=0.9)	Deep Ensemble	0.276
PP MMD (grid)	Moons	Latent (=0.9)	Linear	**0.035**
PP MMD (grid)	Moons	Latent (=0.9)	MLP	**0.002**
PP MMD (grid)	Overlapping	ClapROAR	Deep Ensemble	0.296
PP MMD (grid)	Overlapping	ClapROAR	Linear	0.19
PP MMD (grid)	Overlapping	ClapROAR	MLP	0.374
PP MMD (grid)	Overlapping	Gravitational	Deep Ensemble	0.446
PP MMD (grid)	Overlapping	Gravitational	Linear	0.324
PP MMD (grid)	Overlapping	Gravitational	MLP	0.518
PP MMD (grid)	Overlapping	Latent (=0.5)	Deep Ensemble	0.344
PP MMD (grid)	Overlapping	Latent (=0.5)	Linear	**0.004**
PP MMD (grid)	Overlapping	Latent (=0.5)	MLP	0.49
PP MMD (grid)	Overlapping	Latent (=0.9)	Deep Ensemble	0.362
PP MMD (grid)	Overlapping	Latent (=0.9)	Linear	0.052
PP MMD (grid)	Overlapping	Latent (=0.9)	MLP	0.412

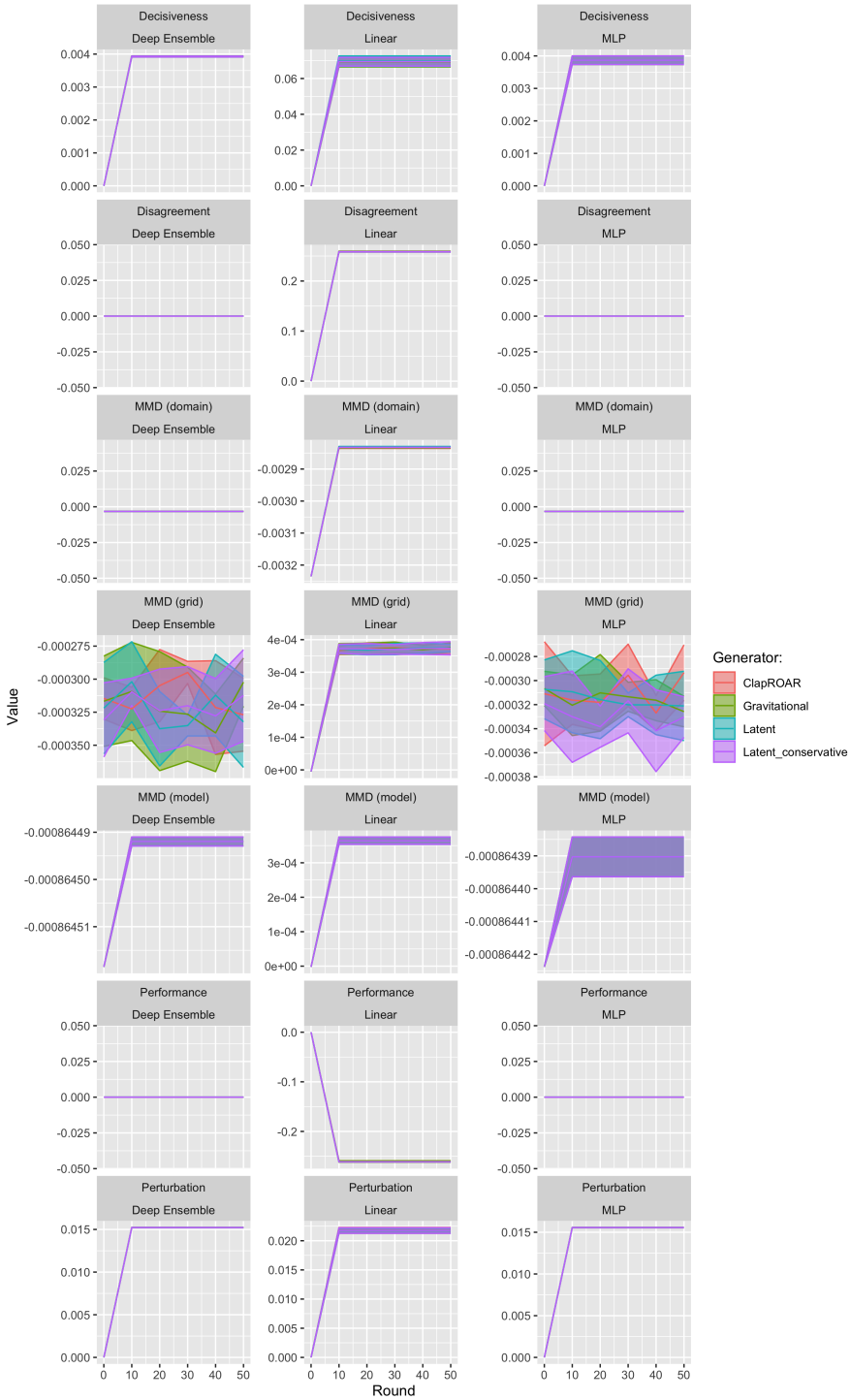


Figure D.29. Evolution of evaluation metrics over the course of the experiment. Data: Circles.

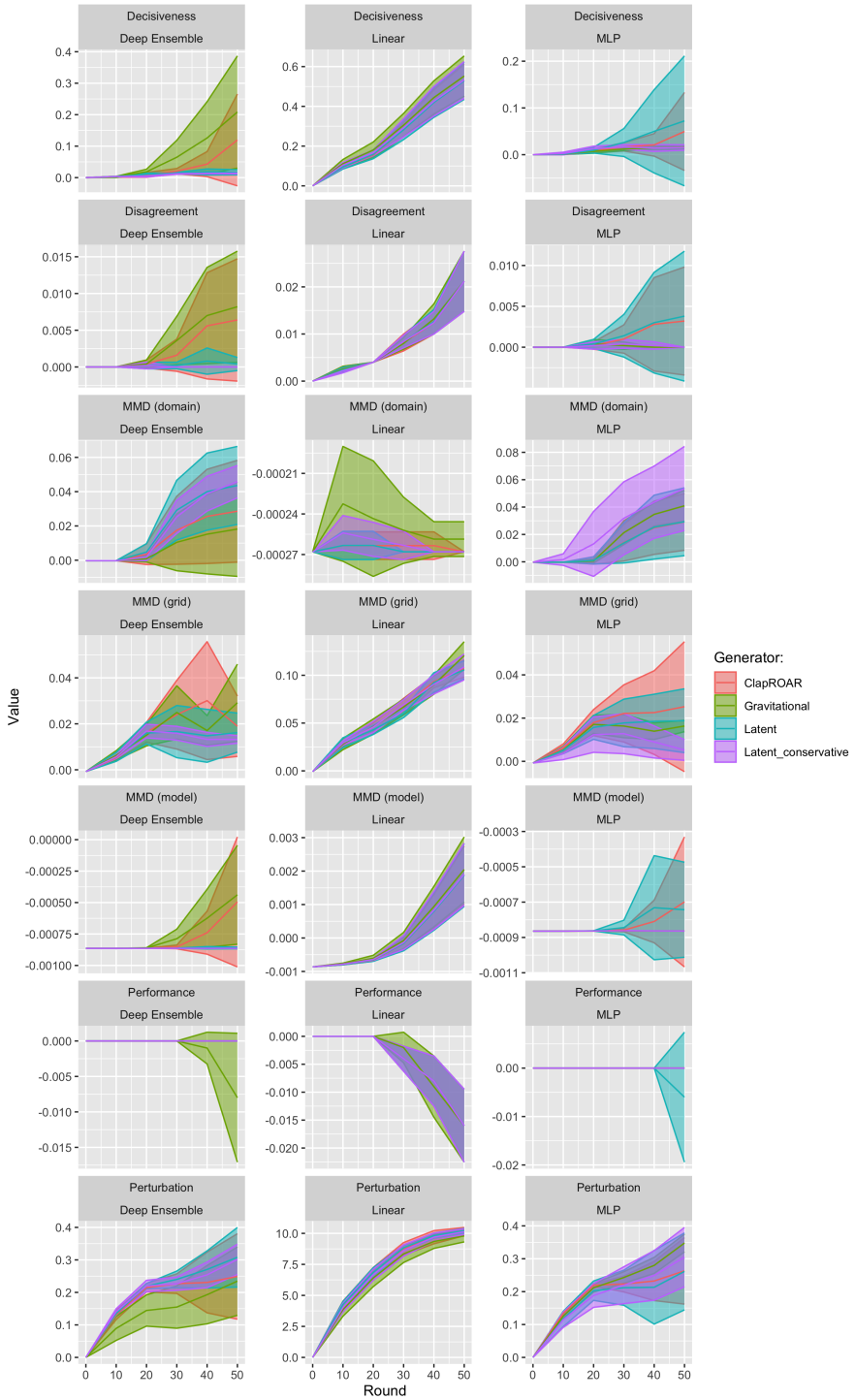


Figure D.30. Evolution of evaluation metrics over the course of the experiment. Data: Linearly Separable.

D

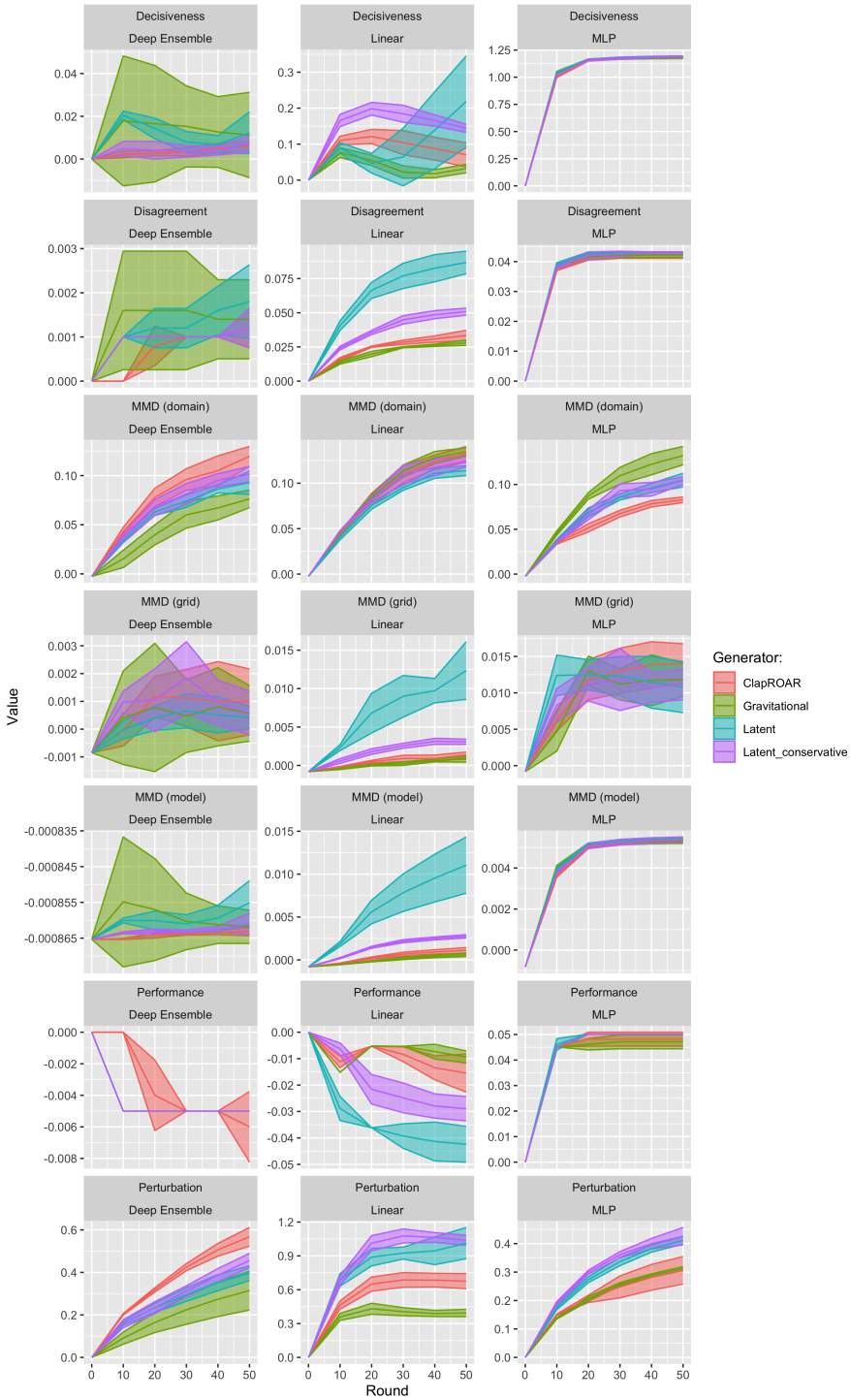


Figure D.31. Evolution of evaluation metrics over the course of the experiment. Data: Moons.

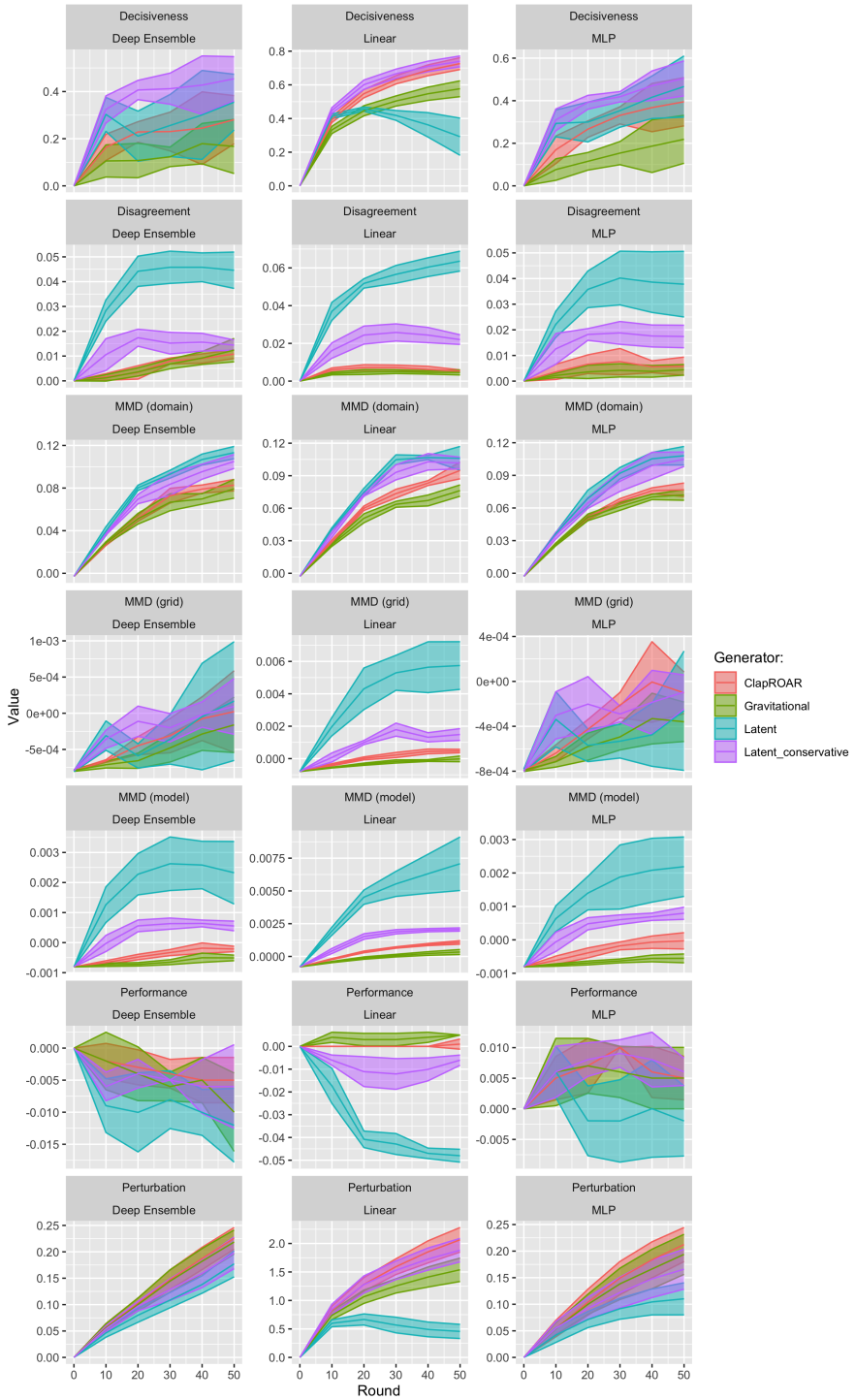


Figure D.32. Evolution of evaluation metrics over the course of the experiment. Data: Overlapping.

D

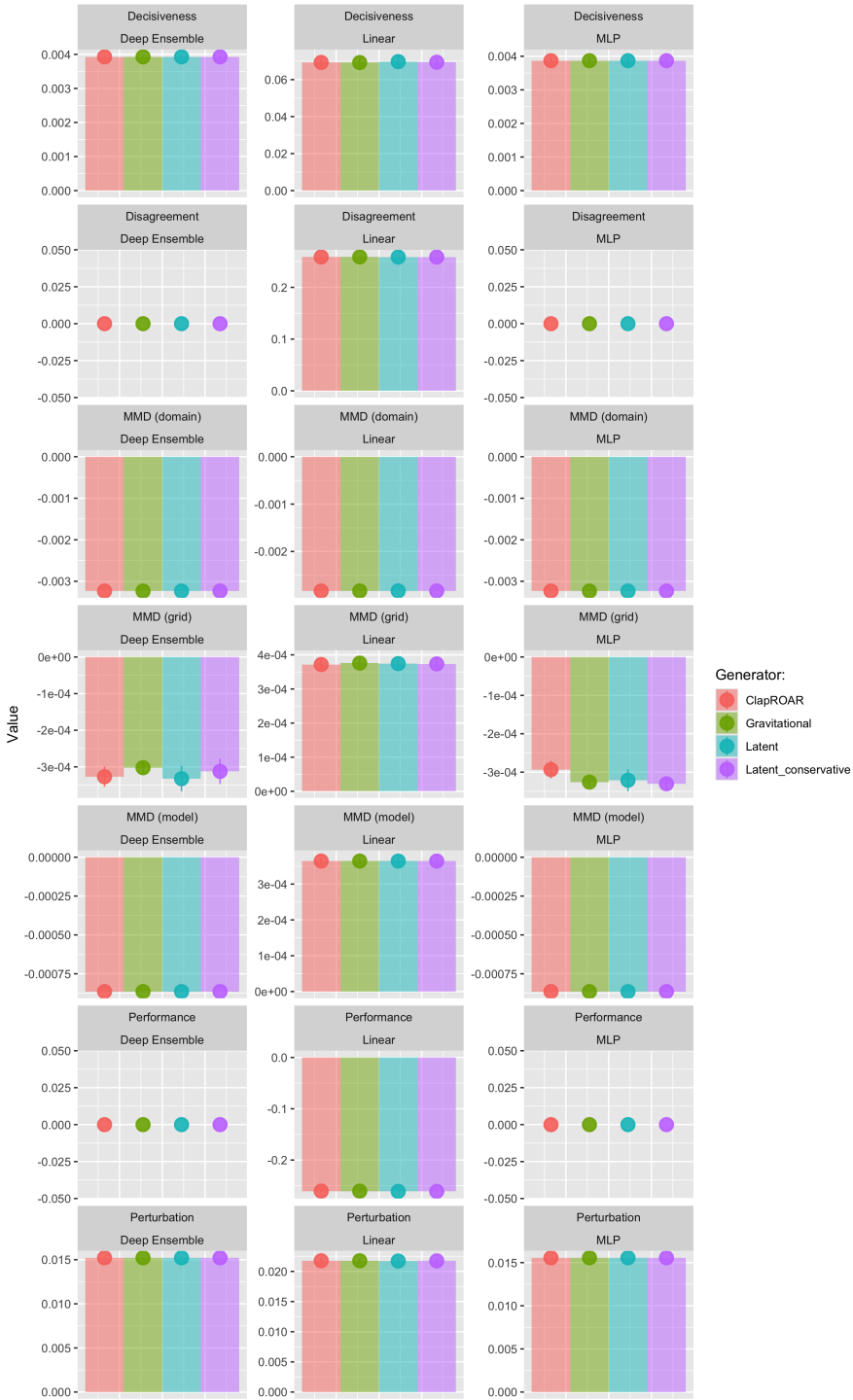


Figure D.33. Evaluation metrics at the end of the experiment. Data: Circles.

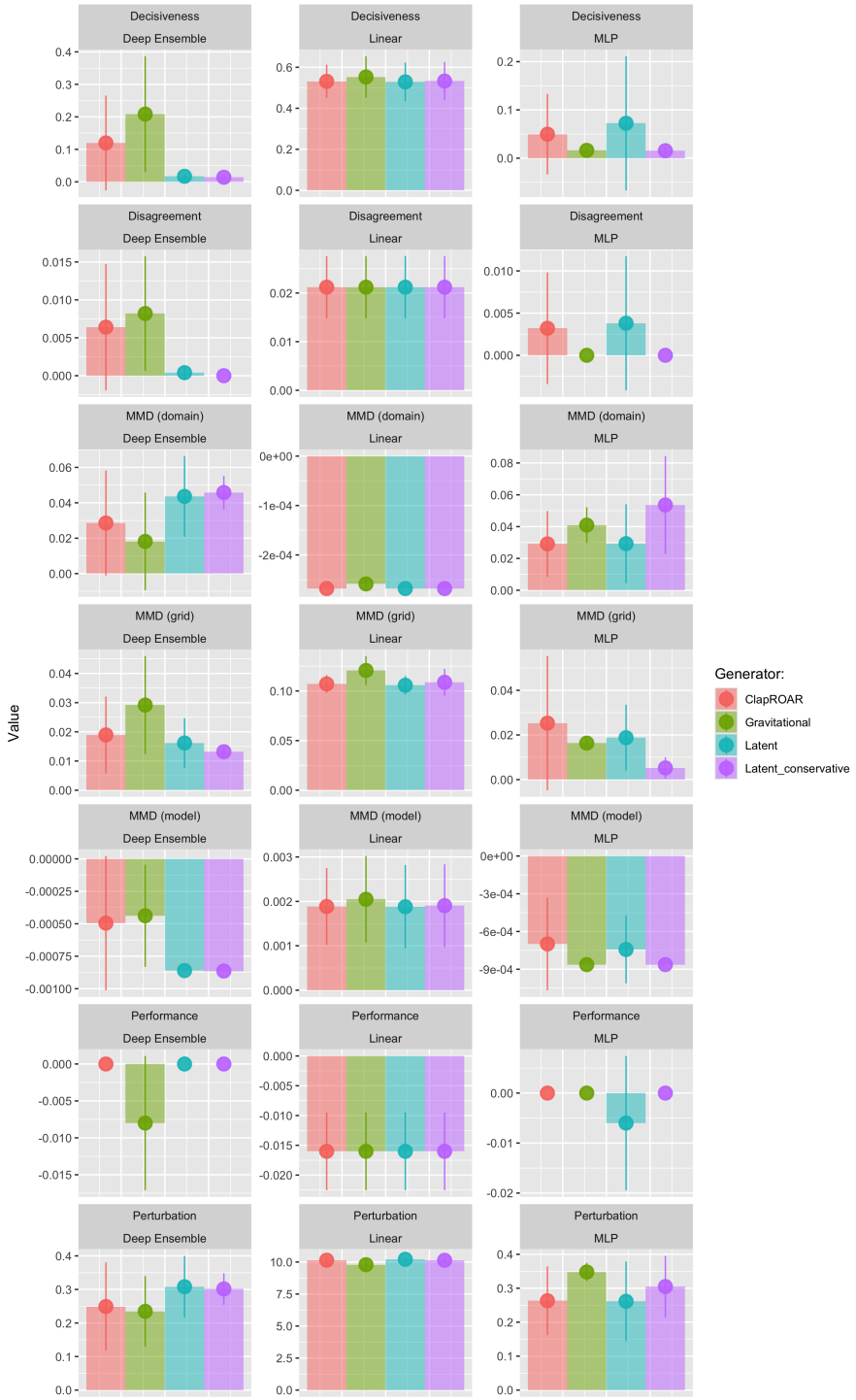


Figure D.34. Evaluation metrics at the end of the experiment. Data: Linearly Separable.

D

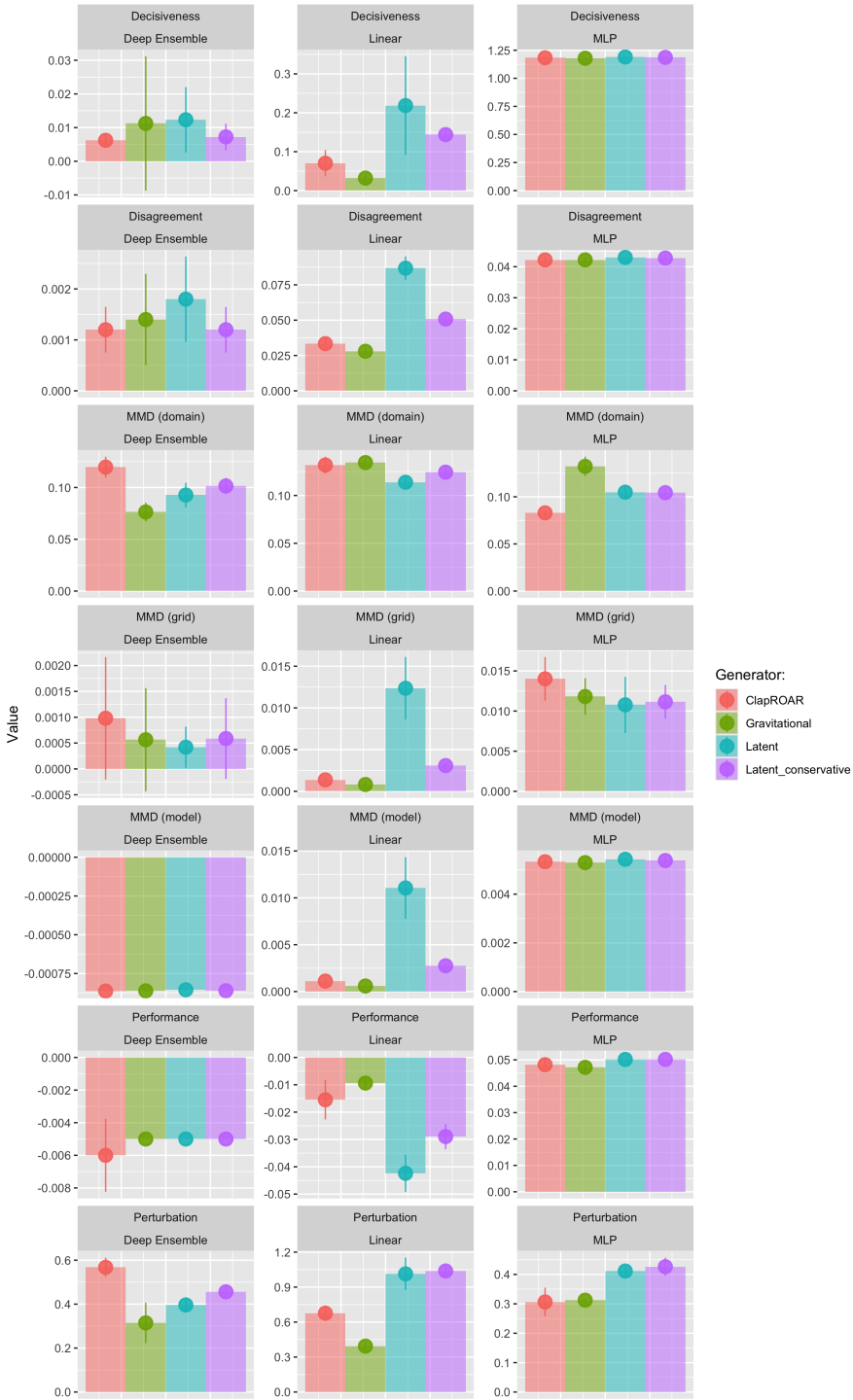


Figure D.35. Evaluation metrics at the end of the experiment. Data: Moons.

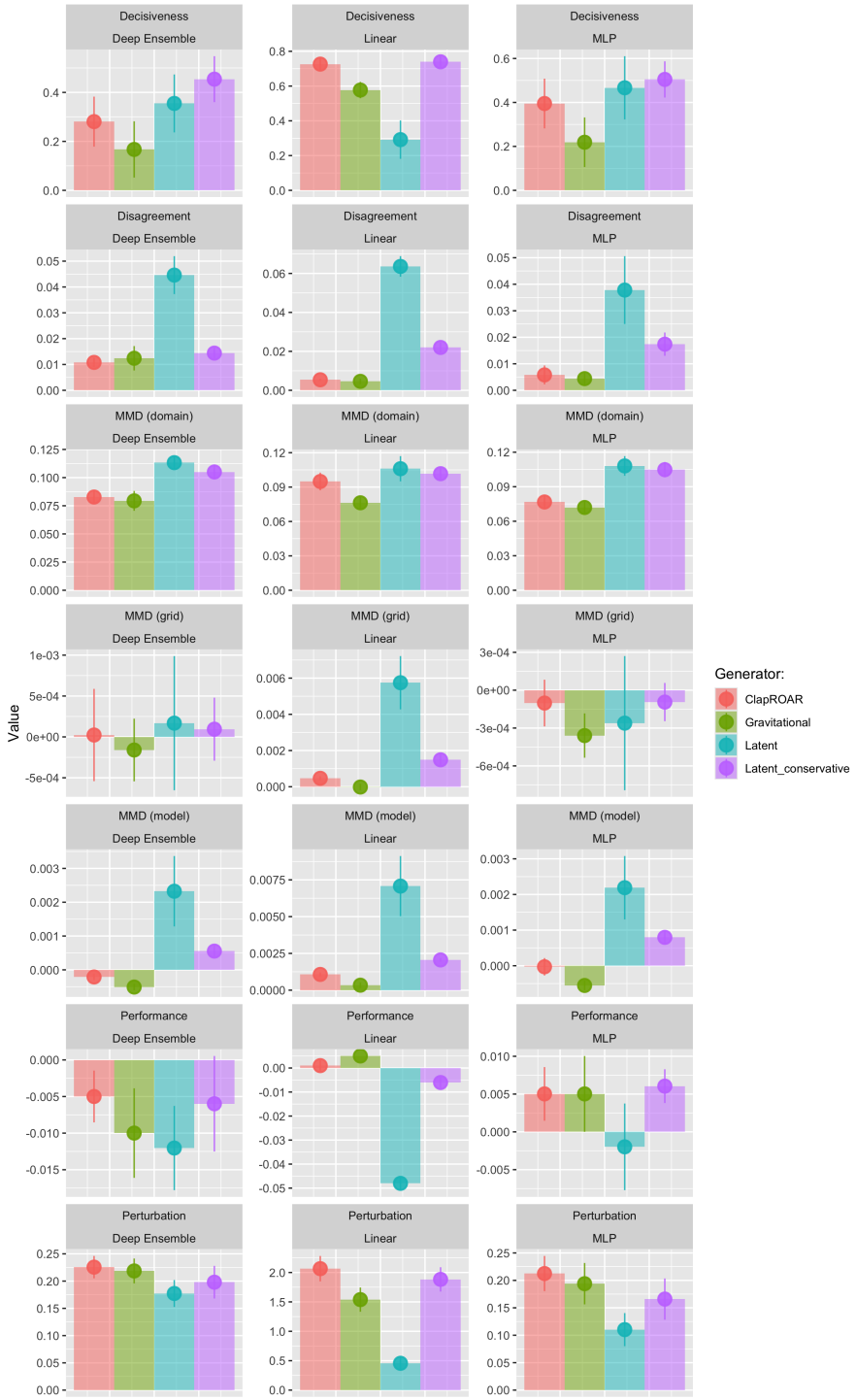


Figure D.36. Evaluation metrics at the end of the experiment. Data: Overlapping.

D

E

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Due to its length, we make the supplementary appendix available separately, instead of including it here. Specifically, the appendix can be found in the preprint of this paper, which has been permanently archived here: https://arxiv.org/pdf/2312.10648.

F

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Due to its length, we make the supplementary appendix available separately, instead of including it here. Specifically, the appendix can be found in the preprint of this paper, which has been permanently archived here: https://arxiv.org/abs/2601.16205.

G

SUPPLEMENTARY MATERIAL FOR CHAPTER 6

In this section, we present additional experimental results that we did not include in the body of the paper for the sake of brevity. We still choose to provide them as additional substantiation of our arguments here. This section also contains additional details concerning the experiment setup for our examples where applicable.

G.1. ARE NEURAL NETWORKS BORN WITH WORLD MAPS?

The initial feature matrix $X^{(n \times m)}$ is made up of $n = 4,217$ and $m = 10$ features. We add a total of 490 random features to X to simulate the fact that not all features ingested by Llama-2 are necessarily correlated with geographical coordinates. That yields 500 features in total. The training subset contains $3,374$ randomly drawn samples, while the remaining 843 are held out for testing. The single hidden layer of the untrained neural network has 400 neurons.

G.2. AUTOENCODERS AS ECONOMIC GROWTH PREDICTORS

This is an additional example that we have not discuss in the body of the paper. Here, we build forth on an application in Economics. However, we now seek to not only predict economic growth from the yield curve, but also extract meaningful features for downstream inference tasks. For this, we will use a neural network architecture.

G.2.1. DATA

To estimate economic growth, we will rely on a quarterly series of the real gross domestic product (GDP) provided by the Federal Reserve Bank of St. Louis. The data arrives in terms of levels of real GDP. In order to estimate growth, we transform the data using log differences. Since our yield curve data is daily, we aggregate it to the quarterly frequency by taking averages of daily yields for each maturity. We also standardize yields since deep learning models tend to perform better with standardized data (Michal S Gal 2019). Since COVID-19 was a substantial structural break in the time series, we also filter out all observations after 2018.

G.2.2. MODEL

Using a simple autoencoder architecture, we let our model g_t denote growth and our conditional $\mathbf{r}_t$ the matrix of aggregated Treasury yield rates at time t. Finally, we let θ denote our model parameters. Formally, we are interested in maximizing the likelihood $p_\theta(g_t|\mathbf{r}_t)$.

The encoder consists of a single fully connected hidden layer with 32 neurons and a hyperbolic tangent activation function. The bottleneck layer connecting the encoder to the decoder, is a fully connected layer with 6 neurons. The decoder consists of two fully connected layers, each with a hyperbolic tangent activation function: the first layer consists of 32 neurons and the second layer will have the same dimension as the input data. The output layer consists of a single neuron for our output variable, g_t. We train the model over 1,000 epochs to minimize mean squared error loss using the Adam optimizer (Kingma and Ba 2017).

The in-sample fit of the model is shown in the left chart of Figure G.1, which shows actual GDP growth and fitted values from the autoencoder model. The model has a large number of free parameters and captures the relationship between economic growth and the yield curve reasonably well, as expected. Since our primary goal is not out-of-sample prediction accuracy but feature extraction for inference, we use all of the available data instead of reserving a hold-out set. As discussed above, we also know that the relationship between economic growth and the yield curve is characterized by two main factors: the level and the spread. Since the model itself is fully characterized by its parameters, we would expect that these two important factors are reflected somewhere in the latent parameter space.

G.2.3. LINEAR PROBE

While the loss function applies most direct pressure on layers near the final output layer, any information useful for the downstream task first needs to pass through the bottleneck layer (Alain and Bengio 2016). On a per-neuron basis, the pressure to distill useful representation is therefore likely maximized there. Consequently, the bottleneck layer activations seem like a natural place to start looking for compact,

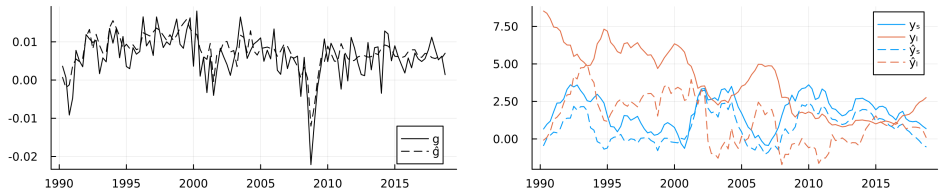


Figure G.1. The left chart shows the actual GDP growth and fitted values from the autoencoder model. The right chart shows the observed average level and spread of the yield curve (solid) along with the predicted values (in-sample) from the linear probe based on the latent embeddings (dashed).

meaningful representations of distilled information. We compute and extract these activations A_t for all time periods $t = 1, ..., T$. Next, we use a linear probe to regress the observed yield curve factors on the latent embeddings. Let Y_t denote the vector containing the two factors of interest in time t: $y_{t,l}$ and $y_{t,s}$ for the level and spread, respectively. Formally, we are interested in the following regression model: $p_w(Y_t|A_t)$ where w denotes the regression parameters. We use Ridge regression with λ set to 0.1. Using the estimated regression parameters $\hat{w}$, we then predict the yield curve factors : $\hat{Y}_t = \hat{w}' A_t$.

The in-sample predictions of the probe are shown in the right chart of Figure G.1. Solid lines show the observed yield curve factors over time, while dashed lines show predicted values. We find that the latent embeddings predict the two yield curve factors reasonably well, in particular the spread.

Did the neural network now learn an intrinsic understanding of the economic relationship between growth and the yield curve? To us, that would be too big of a statement. Still, the current form of information distillation can be useful, even beyond its intended use for monitoring models. For example, an interesting idea could be to use the latent embeddings as features in a more traditional and interpretable econometric model. To demonstrate this, let us consider a simple linear regression model for GDP growth. We might be interested in understanding to what degree economic growth in the past is associated with economic growth today. As we might expect, linearly regressing economic growth on lagged growth, as in column (1) of Table G.1, yields a statistically significant coefficient. However, this coefficient suffers from confounding bias since there are many other confounding variables at play, of which some may be readily observable and measurable, but others may not.

We e.g. already mentioned the relationship between interest rates and economic growth. To account for that, while keeping our regression model as parsimonious as possible, we could include the level and the spread of the US Treasury yield curve as additional regressors. While this slightly changes the estimated magnitude of the coefficient on lagged growth, the coefficients on the observed level and spread are statistically insignificant (column (2) in Table G.1). This indicates that these measures may be too crude to capture valuable information about the relationship

between yields and economic growth. Because we have included two additional regressors with little to no predictive power, the model fit as measured by the Bayes Information Criterium (BIC) has actually deteriorated.

Column (3) of Table G.1 shows the effect of instead including one of the latent embeddings that we recovered above in the regression model. In particular, we pick the one latent embedding that we have found to exhibit the most significant effect on the output variable in a separate regression of growth on all latent embeddings. The estimated coefficient on this latent factor is small in magnitude, but statistically significant. The overall model fit, as measured by the BIC has improved and the magnitude of the coefficient on lagged growth has changed quite a bit. While this is still a very incomplete toy model of economic growth, it appears that the compact latent representation we recovered can be used in order to mitigate confounding bias.

Table G.1. Regression output for various models.

	GDP Growth		
	(1)	(2)	(3)
(Intercept)	0.004***	0.002	0.004***
	(0.001)	(0.002)	(0.001)
Lagged Growth	0.398***	0.385***	0.344***
	(0.087)	(0.089)	(0.088)
Spread		0.000	
		(0.001)	
Level		0.000	
		(0.000)	
Embedding 6			0.008*
			(0.003)
Obs.	114	114	114
BIC	-860.391	-857.429	-864.499
R^2	0.158	0.168	0.203

G

G.3. LLMS FOR ECONOMIC SENTIMENT PREDICTION

G.3.1. LINEAR PROBES

Figure G.2 to Figure G.6 present average performance measures across folds for all indicators each time for the train and test set. We report the correlation between predictions and observed values ('cor'), the mean directional accuracy ('mda'), the mean squared error ('mse') and the root mean squared error ('rmse'). The model depth—as indicated by the number of the layer—increases along the horizontal axis.

Figure G.7 to Figure G.11 present the same performance measures, also for the baseline autoregressive model. Shaded areas show the variation across folds.

G.3.2. SPARK OF ECONONOMIC UNDERSTANDING?

Below we present the 10 sentences in each category that were used to generate the probe predictions plotted in Figure 6.4. In each case, the first 5 sentences were composed by ourselves. The following 5 sentences were generated using ChatGPT 3.5 using the following prompt followed by the examples in each category:

> "I will share 5 example sentences below that sound a bit like they are about price deflation but are really about a deflation in the numbers of doves. Please generate an additional 25 sentences that are similar. Concatenate those sentences to the example string below, each time separating a sentence using a semicolon (just follow the same format I've used for the examples below). Please return only the concatenated sentences, including the original 5 examples.
>
> Here are the examples:"

This was followed up with the following prompt to generate additional sentences:

> "Please generate X more sentences in the same manner and once again return them in the same format. Do not recycle sentences you have already generated, please."

All of the sentences were then passed through the linear probe for the CPI and sorted in ascending or descending order depending on the context (inflation or deflation). We then carefully inspected the list of sentences and manually selected 5 additional sentences to concatenate to the 5 sentences we composed ourselves.

G.3.2.1. INFLATION/PRICES

The following sentences were used:

> Consumer prices are at all-time highs.;Inflation is expected to rise further.;The Fed is expected to raise interest rates to curb inflation.;Excessively loose monetary policy is the cause of the inflation.;It is essential to bring inflation back to target to avoid drifting into hyperinflation territory.;Inflation is becoming a global phenomenon, affecting economies across continents.;Inflation is reshaping the dynamics of international trade and competitiveness.;Inflationary woes are prompting governments to reassess fiscal policies and spending priorities.;Inflation is reshaping the landscape of economic indicators, challenging traditional forecasting models.;The technology sector is not immune to inflation, facing rising costs for materials and talent.

G.3.2.2. INFLATION/BIRDS

The following sentences were used:

> The number of hawks is at all-time highs.;Their levels are expected to rise further.;The Federal Association of Birds is expected to raise barriers of entry for hawks to bring their numbers back down to the target level.;Excessively loose migration policy for hawks is the likely cause of their numbers being so far above target.;It is essential to bring the number of hawks back to target to avoid drifting into hyper-hawk territory.;The unprecedented rise in hawk figures requires a multi-pronged approach to wildlife management.;Environmental agencies are grappling with the task of addressing the inflationary hawk numbers through targeted interventions.;The burgeoning hawk figures highlight the need for adaptive strategies to manage and maintain a healthy avian community.;The unprecedented spike in hawk counts highlights the need for adaptive and sustainable wildlife management practices.;Conservationists advocate for proactive measures to prevent further inflation in hawk numbers, safeguarding the delicate balance of the avian ecosystem.

G.3.2.3. DEFLATION/PRICES

The following sentences were used:

> Consumer prices are at all-time lows.;Inflation is expected to fall further.;The Fed is expected to lower interest rates to boost inflation.;Excessively tight monetary policy is the cause of deflationary pressures.;It is essential to bring inflation back to target to avoid drifting into deflation territory.;The risk of deflation may increase during periods of economic uncertainty.;Deflation can lead to a self-reinforcing cycle of falling prices and reduced economic activity.;The deflationary impact of reduced consumer spending can ripple through the entire economy.;Falling real estate prices can contribute to deflation by reducing household wealth and confidence.;The deflationary impact of falling commodity prices can have ripple effects throughout the global economy.

G.3.2.4. DEFLATION/BIRDS

The following sentences were used:

> The number of doves is at all-time lows.;Their levels are expected to fall further.;The Federal Association of Birds is expected to lower barriers of entry for doves to bring their numbers back up to the target level.;Excessively tight migration policy for doves is the likely cause of

their numbers being so far below target.;Dovelation risks loom large as the number of doves continues to dwindle.;The number of doves is experiencing a significant decrease in recent years.;It is essential to bring the numbers of doves back to target to avoid drifting into dovelation territory.;A comprehensive strategy is needed to reverse the current dove population decline.;Experts warn that without swift intervention, we may witness a sustained decrease in dove numbers.

We think that this sort of manual, LLM-aided adversarial attack against another LLM can potentially be scaled up to allow for rigorous testing, which we will turn to next.

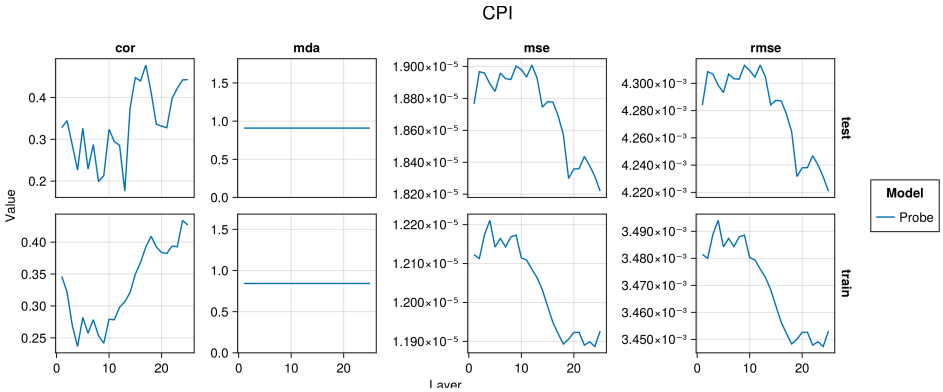


Figure G.2. Average performance measures across folds plotted against model depth (number of layer) for the CPI for the train and test set.

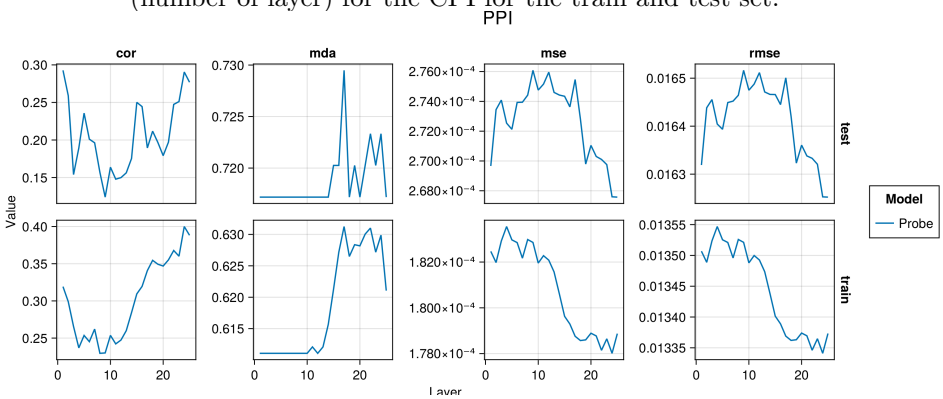


Figure G.3. Average performance measures across folds plotted against model depth (number of layer) for the PPI for the train and test set.

G

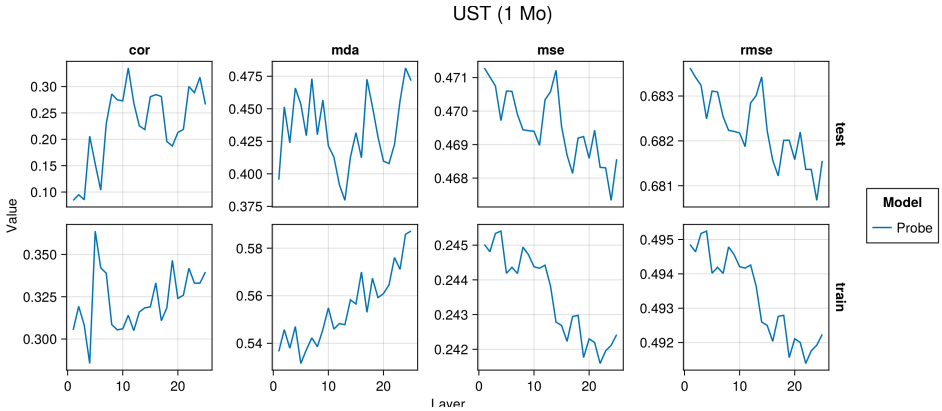


Figure G.4. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Mo) for the train and test set.

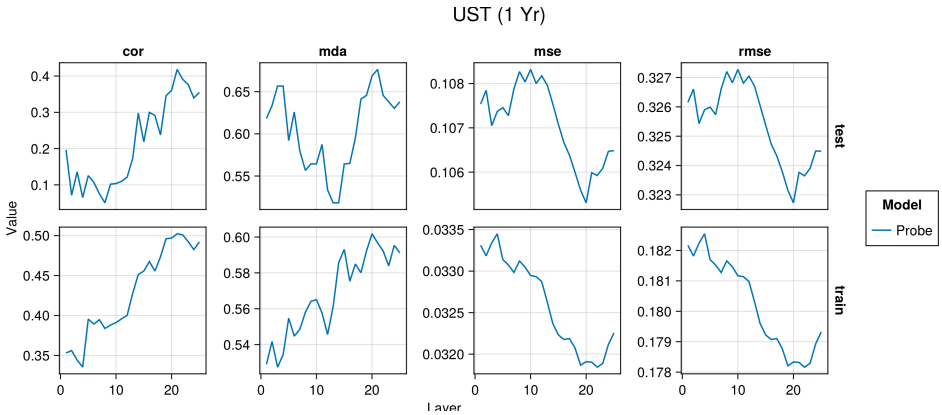


Figure G.5. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Yr) for the train and test set.

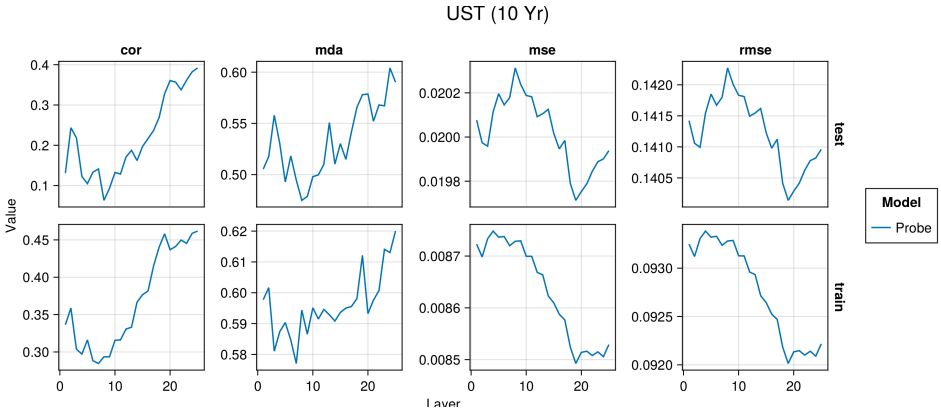


Figure G.6. Average performance measures across folds plotted against model depth (number of layer) for the UST (10 Yr) for the train and test set.

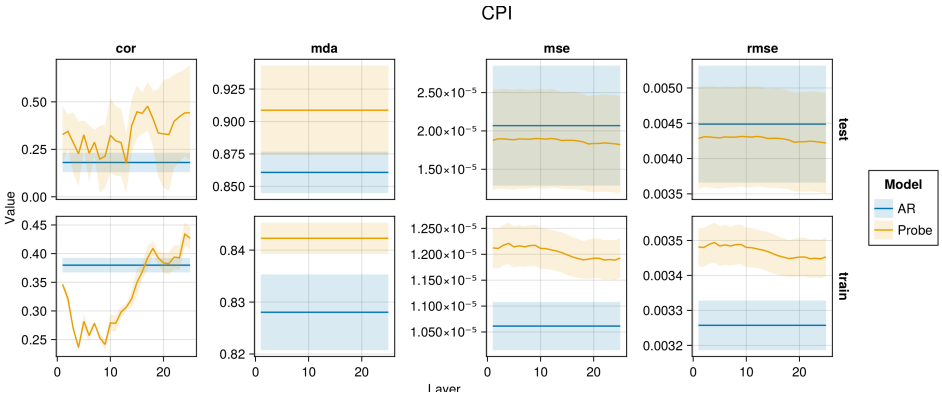


Figure G.7. Average performance measures across folds plotted against model depth (number of layer) for the CPI for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

G

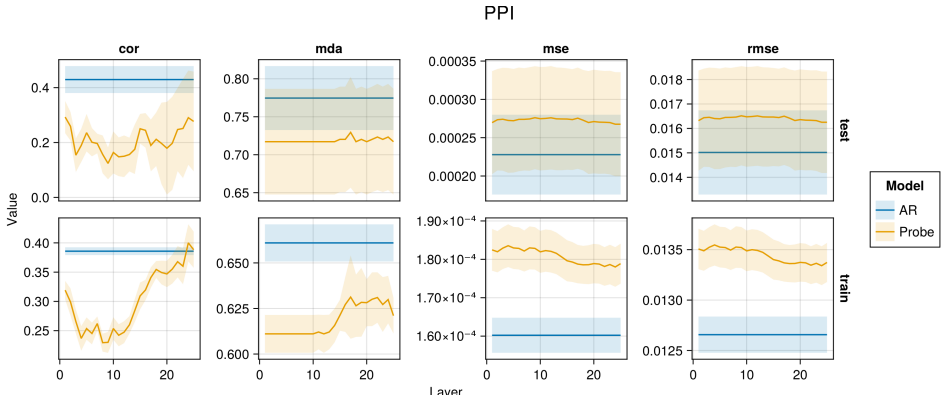


Figure G.8. Average performance measures across folds plotted against model depth (number of layer) for the PPI for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

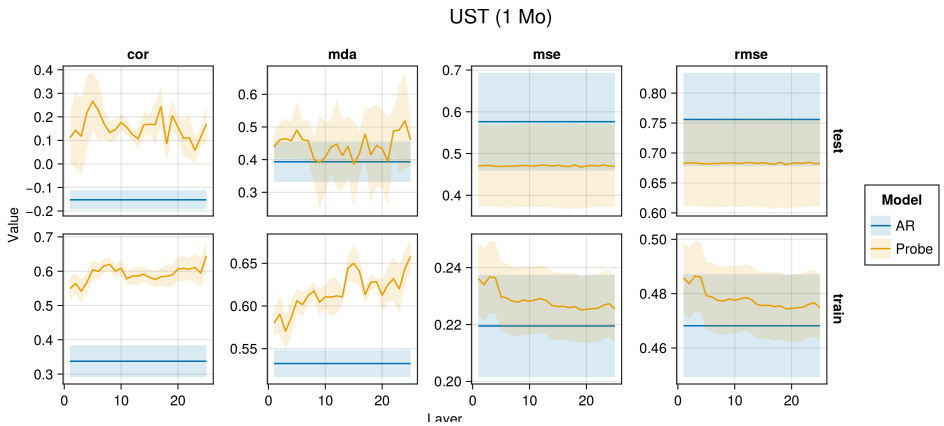


Figure G.9. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Mo) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

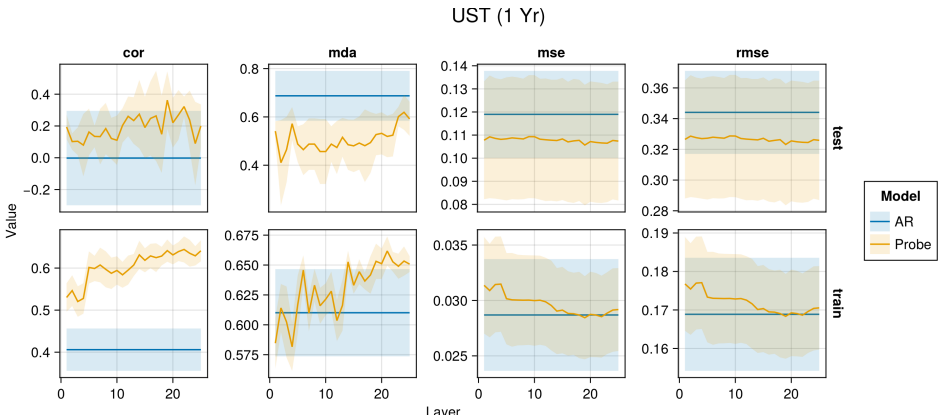


Figure G.10. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Yr) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

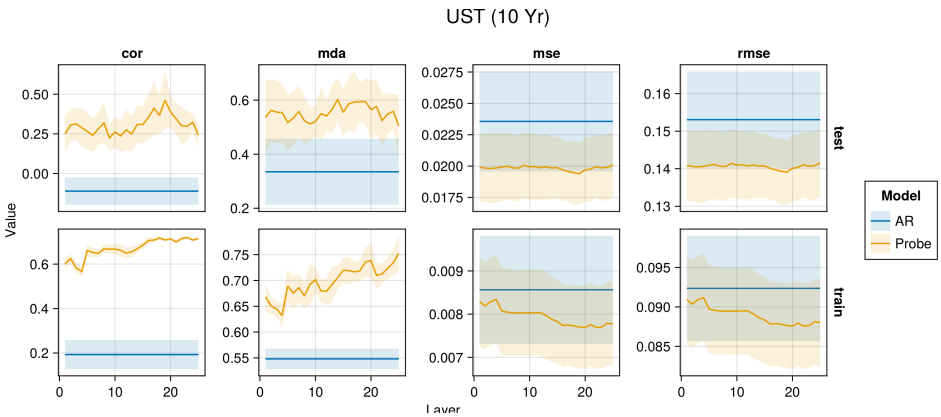


Figure G.11. Average performance measures across folds plotted against model depth (number of layer) for the UST (10 Yr) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

G

G.4. TOWARD PARROT TESTS

In our experiments from Section Section 6.3.3, we considered the following hypothesis tests as a minimum viable testing framework to assess if our probe results (may) provide evidence for an actual 'understanding' of key economic relationships learned purely from text:

Proposition G.1 (Parrot Test)**.**

- H0 (Null)*: The probe never predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$*.*
- H1 (Stochastic Parrots)*: The probe predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$ *for sentences related to the outcome of interest* and *those that are independent (i.e. sentences in all categories).*
- H2 (More than Mere Stochastic Parrots)*: The probe predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$ *for sentences that are related to the outcome variable (IP and DP), but not for sentences that are independent of the outcome (IB and DB).*

To be clear, if in such a test we did find substantial evidence in favour of rejecting both *HO* and *H1*, this would not automatically imply that *H2* is true. But to even continue investigating, if based on having learned meaningful representation the underlying LLM is more than just a parrot, it should be able to pass this simple test.

In this particular case, Figure 6.4 demonstrates that we find some evidence to reject *H0* but not *H1* for *FOMC-RoBERTa*. The median linear probe predictions for sentences about inflation and deflation are indeed substantially higher and lower, respectively than for random noise. Unfortunately, the same is true for sentences about the inflation and deflation in the number of birds, albeit to a somewhat lower degree. This finding holds for both inflation indicators and to a lesser degree also for yields at different maturities, at least qualitatively.

We should note that the number of sentences in each category is very small here (10), so the results in Figure 6.4 cannot be used to establish statistical significance. That being said, even a handful of convincing counter-examples should be enough for us to seriously question the claim, that results from linear probes provide evidence in favor of real 'understanding'. In fact, even a handful of sentences for which any human annotator would easily arrive at the conclusion of independence, a prediction by the probe in either direction casts doubt.

G.5. CODE

All of the experiments were conducted on a MacBook Pro, 14-inch, 2023, with an Apple M2 Pro chip and 16GB of RAM. Forward passes through the FOMC-RoBERTa were run in parallel on 6 threads. All our code will be made publicly available. For the time being, an anonymized version of our code repository can be found here: https://anonymous.4open.science/r/spurious_sentience/README.md.