Development of an Explainable AI Framework for Autonomous Financial Decision-Making

The TRACE Framework: Transparent, Recourse-aware, Accountable, Calibrated and Explainable Decisions

By

Dominic Namachila

Abstract

Financial institutions increasingly delegate decisions about credit, investment and risk to machine-learning systems that act with little or no human intervention. Such autonomy promises speed, consistency and scale, but it collides with a dense set of legal, ethical and practical demands: affected individuals are entitled to reasons, supervisors require that models be validated and governed, and institutions need assurance that a system's behaviour is faithful to the reasons it gives. Post-hoc explanation methods such as SHAP and LIME are widely used to meet these demands, yet they are rarely evaluated against a known truth, and they are seldom integrated with the other mechanisms that an accountable autonomous system requires.

This thesis develops and evaluates TRACE, an explainable AI framework for autonomous financial decision-making. TRACE is organised as four interacting layers: a *decision engine* that produces scores and actions under domain constraints; an *explanation engine* that produces local Shapley attributions, global summaries, decision-level attributions and actionable counterfactual recourse; an *assurance layer* that measures explanation fidelity and stability, audits fairness and proxy discrimination, and uses conformal prediction to decide when the system may act autonomously and when it must escalate to a human; and a *governance layer* that turns these outputs into reason codes, audit records and escalation workflows aligned with credit-reporting, model-risk and AI-regulation requirements.

The framework is evaluated through a programme of controlled computational experiments on synthetic data whose generating process is known exactly. This design makes it possible, unlike with real data, to measure how close explanations are to the true drivers of risk. Seven findings emerge. First, interpretable and constrained models are competitive: a monotonically constrained gradient-boosting model reaches an AUC of 0.845 against 0.850 for the true data-generating model. Second, most of the disagreement between explanations and the true risk drivers is caused by the model rather than by the explainer: KernelSHAP agrees with exact Shapley values of the model at a rank correlation of at least 0.99, whereas the model's own exact Shapley values agree with the ground truth at between 0.87 and 0.96. Third, LIME adds substantial additional error and is less stable under small input perturbations. Fourth, counterfactual recourse that ignores actionability recommends impossible changes in 26% of cases, and most recourse becomes invalid when the model is retrained; a validity margin raises robustness from 37% to 65% at a 43% higher cost. Fifth, marginal conformal prediction meets its overall coverage target while covering only 40% of actual defaulters at the 90% level, a failure that class-conditional calibration corrects at the price of a higher escalation rate. Sixth,

attribution-based auditing detects a proxy variable through which historical label bias enters the model, and removing it reduces the approval-rate gap between groups from 0.166 to 0.116 without loss of accuracy. Seventh, in an autonomous portfolio-allocation agent, the features that dominate forecast explanations are not those that dominate allocation decisions, which demonstrates that explanations must be computed at the level of the decision rather than the prediction.

The thesis contributes a layered reference architecture, a formal account of decision-level attribution and explanation assurance, a ground-truth evaluation methodology for financial XAI, and empirical evidence on the trade-offs between autonomy, accuracy, robustness and accountability. Its limitations, chiefly the reliance on synthetic data and the absence of a study with human users, are stated explicitly and define an agenda for future work.

Contents

Abstract i

List of Abbreviations ix

1 Introduction 1

1.1 Context and motivation 1
1.2 Problem statement 2
1.3 Aim and research questions 3
1.4 Research approach 4
1.5 Contributions 4
1.6 Scope 5
1.7 Structure of the thesis 5

2 Background and Literature Review 6

2.1 Machine learning in financial decision-making 6
2.1.1 Credit risk 6
2.1.2 Portfolio management and trading 7
2.1.3 Autonomy and automation 7
2.2 Explainable artificial intelligence 7
2.2.1 Concepts and desiderata 7
2.2.2 Feature-attribution methods 8
2.2.3 Explainability in finance 9
2.3 Evaluating explanations 9
2.4 Counterfactual explanations and algorithmic recourse 10
2.5 Fairness and proxy discrimination 11
2.6 Uncertainty quantification and selective decision-making 11
2.7 Governance, accountability and documentation 12
2.8 Regulatory landscape 12
2.9 Synthesis and research gaps 12

3 Research Methodology 14

3.1 Research paradigm 14
3.2 Requirements derivation 14
3.3 Evaluation strategy 14
3.3.1 Why synthetic data 14
3.3.2 Threats arising from the choice 15

3.3.3 Evaluation criteria 16
3.4 Ground-truth attribution 16
3.5 Software and reproducibility 17
3.6 Ethical considerations 17

4 The TRACE Framework 18
4.1 Design principles 18
4.2 Architecture overview 18
4.3 Layer 1: Decision engine 19
4.3.1 Constrained scoring models 19
4.3.2 Decision rules 20
4.4 Layer 2: Explanation engine 20
4.4.1 Local attribution 20
4.4.2 Reason codes 20
4.4.3 Actionable and robust recourse 21
4.4.4 Decision-level attribution 21
4.5 Layer 3: Assurance layer 22
4.5.1 Fidelity and stability monitoring 22
4.5.2 Conformal autonomy gate 23
4.5.3 Fairness and proxy audit 23
4.6 Layer 4: Governance layer 24
4.7 Summary 24

5 Experimental Environment and Protocol 26
5.1 Synthetic credit environment 26
5.1.1 Features 26
5.1.2 True risk function 26
5.1.3 Biased environment 27
5.1.4 Data splits 27
5.2 Models 27
5.3 Explanation methods 28
5.4 Metrics 28
5.5 Experiment protocols 29
5.6 Synthetic market environment 30

6 Results I: Autonomous Credit Decisions 31
6.1 E1: Predictive performance of constrained and flexible models 31
6.2 E2: Explanation fidelity against ground truth 32
6.2.1 Total fidelity 32
6.2.2 Decomposing the error: model gap versus explainer gap 32

6.2.3 Sensitivity to the background sample ... 34
6.3 E3: Explanation stability ... 34
6.4 E4: Actionable and robust recourse ... 35
6.5 E5: The conformal autonomy gate ... 36
6.6 E6: Detecting proxy discrimination with explanations ... 37
6.7 An end-to-end decision ... 38
6.8 Summary ... 39

7 Results II: Autonomous Portfolio Allocation 42
7.1 Investment performance ... 42
7.2 Forecast-level explanations ... 43
7.3 Decision-level explanations ... 43
7.4 Explaining a single rebalancing decision ... 44
7.5 Summary ... 45

8 Discussion 46
8.1 Answers to the research questions ... 46
8.2 Implications for practice ... 47
8.3 Implications for regulation and supervision ... 48
8.4 Relation to the literature ... 48
8.5 Threats to validity ... 49
8.5.1 External validity ... 49
8.5.2 Internal validity ... 49
8.5.3 Construct validity ... 50
8.5.4 Statistical conclusion validity ... 50
8.6 Ethical reflections ... 50

9 Conclusion and Future Work 51
9.1 Summary of the thesis ... 51
9.2 Principal findings ... 51
9.3 Contributions ... 52
9.4 Limitations ... 52
9.5 Future work ... 53
9.6 Closing remarks ... 53

References 55

A Data-Generating Process 60
A.1 Credit environment ... 60
A.2 Market environment ... 61

B Implementation Details **62**

C Audit-Record Schema **63**

List of Figures

4.1 The TRACE architecture: four layers and the principal information flows. 19

6.1 E2: Agreement of explanations with the ground-truth attribution, by model and explainer. 32
6.2 E2: Model gap compared with the explainer gaps of KernelSHAP and LIME. 33
6.3 E2: Mean absolute attribution by feature for the ground truth, the GBM and the monotone HGB. 34
6.4 E4: Actionability, cost and robustness of recourse. 35
6.5 E5: Class-conditional coverage (left) and escalation rate (right) for marginal and Mondrian conformal prediction 36
6.6 E6: Feature importance (left) and correlation of each feature's attribution with group membership (right), for models trained on clean and on biased labels. 37
6.7 Explanation of one declined applicant: contributions of each feature to the log-odds of default, from the population base value to the applicant's score 39

7.1 E7: Cumulative wealth of the strategies, net of transaction costs. 43
7.2 E7: Left, share of forecast attribution by feature at each annual retraining. Right, share of decision-level attribution by signal family, averaged over all weekly decisions 44
7.3 E7: Decision-level attribution for one weekly rebalancing. Bars show the contribution of each signal family to the change in each asset's weight relative to the reference portfolio; diamonds show the total change. 45

List of Tables

3.1 Requirements for an explainable autonomous financial decision-making system. 15

4.1 Mapping of TRACE components to requirements and obligations. 24

6.1 E1: Test-set predictive performance (n = 10,000). 31

6.2 E2: Fidelity of explanations to the ground-truth attribution (200 applicants). Rank correlation with bootstrap 95% CI; relative ℓ_2 error is not reported for LIME, which explains probabilities rather than log-odds. . . 40

6.3 E2: Decomposition of explanation error. The model gap is 1− rank correlation between the model's exact attribution and the ground truth; the other columns give rank correlation between each explainer and the model's exact attribution. 40

6.4 E3: Stability of explanations (monotone HGB). 40

6.5 E4: Recourse quality for 150 declined applicants. 40

6.6 E5: Conformal autonomy gate on the test set. Coverage overall, for defaulters (y = 1), non-defaulters (y = 0) and by group; escalation rate; accuracy of autonomous decisions. 41

6.7 E6: Group outcomes and proxy diagnostics. DP: demographic-parity gap in approval; EO: approval gap among truly low-risk applicants; Imp. and Corr.: mean absolute attribution of postcode_index and its correlation with group; AUC against outcomes drawn from the clean risk function. . 41

6.8 Audit-record excerpt for the declined applicant: values, recourse and exact Shapley attribution (log-odds). 41

7.1 E7: Out-of-sample performance over 787 weeks, net of 10 bp transaction costs. Returns, volatility and drawdown in per cent; turnover is the mean weekly sum of absolute weight changes. 42

B.1 Hyperparameters and settings. 62

List of Abbreviations

AUC	Area under the receiver operating characteristic curve
CF	Counterfactual
DP	Demographic parity
DTI	Debt-to-income ratio
ECE	Expected calibration error
EO	Equal opportunity
GBM	Gradient boosting machine
HGB	Histogram-based gradient boosting
LIME	Local Interpretable Model-agnostic Explanations
MAD	Median absolute deviation
MLP	Multilayer perceptron
MV	Mean–variance
SHAP	SHapley Additive exPlanations
TRACE	Transparent, Recourse-aware, Accountable, Calibrated, Explainable
XAI	Explainable artificial intelligence

Chapter 1

Introduction

1.1 Context and motivation

Finance was one of the first industries to automate judgement. Credit scoring replaced the loan officer's discretion with statistical scorecards decades before the term "artificial intelligence" entered everyday business language, and algorithmic execution replaced much of the trading floor. What has changed in the last decade is not the existence of automated financial decisions but their character. Contemporary systems are built from flexible machine-learning models such as gradient-boosted tree ensembles and deep neural networks; they consume more data, adapt more quickly and act with less human mediation than their predecessors. A mortgage pre-approval, a credit-limit increase, a fraud block on a card transaction or a portfolio rebalance may now be executed end-to-end by software, with a human reviewing only a sample of outcomes or only the exceptions that the software itself chooses to escalate. This thesis refers to such systems as *autonomous financial decision-making systems*.

Autonomy brings real benefits. Automated decisions are fast, which matters to a customer waiting for a loan and to a portfolio exposed to a moving market. They are consistent, in the sense that identical inputs receive identical treatment, which removes some forms of arbitrary human variation. They scale at negligible marginal cost, which can widen access to financial services for people whose applications would not justify the expense of manual underwriting. And flexible models frequently extract more predictive signal from data than the linear scorecards they replace (Lessmann et al., 2015).

Autonomy also concentrates risk. A flawed model applied to millions of decisions produces errors at the same scale. A model that has learned a proxy for a protected characteristic can discriminate systematically and invisibly. A model that is accurate on average may be badly wrong for a sub-population, or may fail abruptly when economic conditions shift. And when an autonomous system declines an application or sells an asset, someone must be able to say why. The person affected has a moral and often a legal claim to an explanation; the institution needs to understand its own exposure; and supervisors need evidence that the system is under control.

These concerns have been codified. In the United States, the Equal Credit Opportunity Act and its implementing Regulation B require creditors to give applicants the specific principal reasons for adverse action, and supervisory guidance on model risk

management (SR 11-7) requires that models be conceptually sound, validated and monitored (Board of Governors of the Federal Reserve System, 2011). In the European Union, the General Data Protection Regulation restricts solely automated decisions with legal or similarly significant effects and is widely read as requiring meaningful information about the logic involved (Goodman and Flaxman, 2017; Wachter et al., 2017). The EU Artificial Intelligence Act, adopted in 2024, classifies AI systems used to evaluate the creditworthiness of natural persons as high-risk and imposes obligations concerning transparency, human oversight, accuracy, robustness and record-keeping (European Parliament and Council, 2024). Across jurisdictions the direction of travel is the same: autonomous financial decisions are permitted, but only if they can be explained, overseen and audited.

The field of explainable artificial intelligence (XAI) has responded with a large toolkit (Adadi and Berrada, 2018; Guidotti et al., 2018; Arrieta et al., 2020). Feature-attribution methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) assign each input a share of responsibility for a prediction; counterfactual methods describe the smallest change that would have altered the outcome (Wachter et al., 2017; Mothilal et al., 2020); global methods such as partial-dependence and accumulated-local-effects plots summarise model behaviour (Friedman, 2001; Apley and Zhu, 2020); and intrinsically interpretable models avoid the need for post-hoc explanation altogether (Rudin, 2019). In financial practice, SHAP in particular has become a de facto standard for generating reason codes from tree ensembles (Bussmann et al., 2021).

1.2 Problem statement

Despite this toolkit, three gaps separate the state of the art from what accountable autonomous financial systems require.

The verification gap. Explanations are almost never evaluated against the truth. On real data the true causes of default or of asset returns are unknown, so explanations are assessed by proxies: agreement between methods, plausibility to experts, or the change in model output when "important" features are removed (Hooker et al., 2019). None of these tells an institution whether the reasons it gives a customer are the reasons that actually drive that customer's risk, nor whether discrepancies stem from the explainer or from the model. Explanation methods have also been shown to be unstable (Alvarez-Melis and Jaakkola, 2018) and manipulable (Slack et al., 2020), which makes unverified explanations a weak foundation for compliance.

The integration gap. Explanation is usually treated as an add-on to a predictive model, separate from the other properties an autonomous system needs. Yet in practice

these properties are intertwined. Whether a decline can be explained with actionable recourse depends on which features the model uses and how it was constrained. Whether a decision should be taken autonomously at all depends on the uncertainty of the prediction. Whether an explanation is trustworthy depends on whether it is stable and faithful. Whether the system is fair can often be diagnosed through its explanations. A framework is needed that treats these concerns as components of one architecture rather than as unrelated tools.

The decision gap. Most XAI research explains *predictions*: a probability of default, a forecast return. Autonomous systems, however, take *decisions*: approve or decline, buy or sell, at what weight. Between prediction and decision lie thresholds, optimisers, constraints and interactions among many predictions. An explanation of a forecast does not necessarily explain the action that the system took, and as this thesis demonstrates, it can point to quite different drivers.

1.3 Aim and research questions

The aim of this research is to develop, formalise and evaluate an explainable AI framework that enables financial decisions to be taken autonomously while remaining transparent, contestable, reliable and auditable. The research is structured around five questions.

RQ1 What architectural components and interfaces does an explainable autonomous financial decision-making system require, and how should they interact?

RQ2 How faithful are widely used post-hoc explanation methods to the true drivers of financial risk, and how much of their error is attributable to the explainer as opposed to the model being explained?

RQ3 How can explanations be made actionable and robust for the individuals affected by autonomous decisions, in particular through counterfactual recourse that respects real-world constraints and survives model updates?

RQ4 How can uncertainty quantification determine when an autonomous system may act and when it must escalate a decision to a human, and with what guarantees for different groups of cases?

RQ5 How can explanations be lifted from the level of individual predictions to the level of decisions taken by an optimiser, and do the two levels differ in practice?

1.4 Research approach

The research follows a design-science approach (Hevner et al., 2004; Peffers et al., 2007). An artefact, the TRACE framework, is designed from requirements derived from regulation, supervisory practice and the XAI literature; it is instantiated in software; and it is evaluated through controlled computational experiments. The central methodological choice is to evaluate on synthetic data generated from a fully specified process. Because the true risk function is known, it is possible to compute the ground-truth Shapley attribution of every applicant's risk and to compare explanations with it. Synthetic data also allows historical bias to be injected in a controlled way, so that the ability of explanations to reveal proxy discrimination can be tested against a known answer. The cost of this choice is external validity, which the thesis addresses explicitly in Chapter 8.

1.5 Contributions

The thesis makes the following contributions.

1. **A reference architecture.** TRACE, a four-layer framework comprising a decision engine, an explanation engine, an assurance layer and a governance layer, with defined interfaces, data contracts and an audit-record schema (Chapter 4).
2. **Formal foundations for decision-level explanation and explanation assurance.** A formulation of Shapley attribution over feature groups for optimisation-based decisions, a decomposition of explanation error into a model gap and an explainer gap, and an integration of conformal prediction as an autonomy gate (Chapter 4).
3. **A ground-truth evaluation methodology.** A synthetic credit environment with a known, non-linear, interacting risk function and a controllable bias mechanism, together with exact interventional Shapley computation for both the true function and fitted models (Chapters 3 and 5).
4. **Empirical evidence** on explanation fidelity and stability, actionable and robust recourse, conformal autonomy gating, proxy-discrimination auditing and decision-level attribution in a portfolio agent (Chapters 6 and 7).
5. **Design guidance** for practitioners and supervisors, mapping framework outputs to regulatory obligations and identifying where current practice is likely to mislead (Chapter 8).

1.6 Scope

The thesis concentrates on two representative decision types: consumer credit approval, which is a high-volume decision about individuals with strong legal duties of explanation, and portfolio allocation, which is a sequential decision produced by an optimiser acting on many forecasts. Fraud detection, insurance pricing and market-making share many of the issues discussed but are not studied directly. The thesis addresses tabular data and does not treat explanation of large language models or text-based decision systems. It does not include experiments with human participants; questions about how people perceive and use explanations are discussed but left to future work.

1.7 Structure of the thesis

Chapter 2 reviews the literature on automated financial decision-making, explanation methods, their evaluation, recourse, fairness, uncertainty quantification and regulation, and identifies the gaps that motivate the research. Chapter 3 sets out the research methodology. Chapter 4 presents the TRACE framework and its formal foundations. Chapter 5 describes the experimental environment and protocol. Chapter 6 reports results for autonomous credit decisions and Chapter 7 for autonomous portfolio allocation. Chapter 8 discusses implications, regulatory alignment and threats to validity. Chapter 9 concludes and sets out future work. Appendices specify the data-generating process, hyperparameters and the audit-record schema in full.

Chapter 2

Background and Literature Review

This chapter reviews the bodies of work on which the thesis builds. It begins with machine learning for financial decisions, turns to the concepts and methods of explainable AI, examines how explanations are evaluated, and then reviews counterfactual recourse, fairness, uncertainty quantification and human oversight. It closes with the regulatory landscape and a synthesis of the research gaps.

2.1 Machine learning in financial decision-making

2.1.1 Credit risk

Credit scoring is the canonical application of statistical learning in finance. Logistic-regression scorecards, often built on binned and weight-of-evidence-transformed variables, dominated industrial practice for decades because they are stable, easy to validate and directly yield reason codes. A substantial literature has since compared them with flexible learners. In a large benchmark of 41 classifiers on eight credit data sets, Lessmann et al. (2015) found that ensembles, particularly random forests and heterogeneous ensembles, systematically outperformed logistic regression, though often by modest margins. Khandani et al. (2010) showed that machine-learning models built on transaction data could forecast consumer delinquencies with economically significant improvements, and systematic reviews (Dastile et al., 2020) confirm that tree ensembles and neural networks are now the principal objects of study. Gradient boosting, especially in efficient implementations such as XGBoost (Chen and Guestrin, 2016), has become the workhorse of industrial credit modelling.

The gains from flexibility are not distributed evenly. Fuster et al. (2022) found that moving from logistic to tree-based models in US mortgage lending would increase predicted access to credit overall but that the benefits would accrue unequally across racial and ethnic groups, because flexible models exploit correlations among variables that are themselves associated with group membership. Bartlett et al. (2022) documented that both face-to-face and algorithmic lenders charged Latinx and African-American borrowers higher rates, although algorithmic lenders discriminated less in approval decisions. These studies show that the move to flexible models raises distributional and legal questions, not only technical ones.

2.1.2 Portfolio management and trading

Portfolio theory begins with Markowitz (1952), whose mean–variance framework chooses weights to balance expected return against variance. The practical weakness of mean–variance optimisation is its sensitivity to estimation error in expected returns; DeMiguel et al. (2009) showed that an equally weighted portfolio is hard to beat out of sample when means are estimated from history. Two responses are relevant here. The first improves covariance estimation through shrinkage (Ledoit and Wolf, 2004). The second improves expected-return estimation through machine learning: Gu et al. (2020) showed that trees and neural networks substantially improve the out-of-sample prediction of equity returns by exploiting non-linear interactions among characteristics. A parallel stream learns allocation policies end to end, for instance by direct reinforcement (Moody and Saffell, 2001) or deep reinforcement learning (Jiang et al., 2017); Heaton et al. (2017) and Dixon et al. (2020) survey deep learning in this setting. In all of these approaches, the decision (a vector of weights) is the output of an optimiser or policy acting on many predictions, which is the structural fact that motivates the decision-level explanation developed in Chapter 4.

2.1.3 Autonomy and automation

Automation is not binary. Parasuraman et al. (2000) distinguish levels of automation ranging from systems that merely offer suggestions to systems that act and inform the human only if asked, or not at all, and they emphasise that the appropriate level depends on the consequences of error and the reliability of the system. In financial services, the most common design is *exception-based* autonomy: routine cases are decided by the model, and cases meeting certain conditions are referred to a human. The literature offers little guidance on how the referral conditions should be set so as to guarantee properties of the autonomous decisions, which is the problem the autonomy gate in TRACE addresses.

2.2 Explainable artificial intelligence

2.2.1 Concepts and desiderata

There is no single agreed definition of interpretability or explainability. Doshi-Velez and Kim (2017) define interpretability as the ability to explain or present in understandable terms to a human and call for rigorous evaluation. Lipton (2018) warns that the term is used for heterogeneous properties, including simulatability, decomposability and post-hoc interpretability, and that these are not interchangeable. Miller (2019), drawing

on philosophy and psychology, argues that human explanations are contrastive ("why P rather than Q"), selective (a few causes, not all), and social, which favours counterfactual and sparse explanations. Surveys by Adadi and Berrada (2018), Guidotti et al. (2018), Arrieta et al. (2020) and the textbook of Molnar (2022) organise methods along common axes: intrinsic versus post-hoc, model-specific versus model-agnostic, and local versus global.

A recurring theme is the distinction between explanations that are *faithful*, meaning they reflect what the model actually computes, and those that are merely *plausible* to a human. Rudin (2019) argues that for high-stakes decisions the gap between the two is dangerous and that inherently interpretable models should be preferred where they are competitive. This thesis takes the position that the choice is empirical: where a constrained, more interpretable model loses little accuracy it should be preferred, and where post-hoc explanations are used their fidelity should be measured rather than assumed.

2.2.2 Feature-attribution methods

LIME. Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016) fits a sparse linear surrogate to the model's outputs on random perturbations of the instance, weighted by proximity. The surrogate's coefficients are read as the explanation. LIME is model-agnostic and fast, but its output depends on the perturbation distribution, the kernel width and the random seed, and it carries no guarantee that the surrogate's coefficients are additive contributions to the specific prediction.

Shapley values and SHAP. The Shapley value (Shapley, 1953) is the unique allocation of a cooperative game's payoff among players satisfying efficiency, symmetry, dummy and additivity. Štrumbelj and Kononenko (2014) and Lundberg and Lee (2017) applied it to prediction by treating features as players and defining the value of a coalition S as the expected model output when the features in S are fixed at the instance's values. Lundberg and Lee (2017) unified several earlier methods under the SHAP framework and introduced KernelSHAP, a weighted linear regression estimator of Shapley values. TreeSHAP (Lundberg et al., 2020) computes exact Shapley values for tree ensembles in polynomial time, which explains SHAP's dominance in finance.

The definition of the coalition value is contested. The *observational* (conditional) version conditions on the features in S under the data distribution; the *interventional* (marginal) version replaces the absent features with draws from their marginal distribution. Janzing et al. (2020) argue on causal grounds that the interventional version correctly attributes the effect of the model's inputs on its output, while Aas et al. (2021) develop more accurate conditional estimators for dependent features, and Chen et al.

(2020) frame the choice as being "true to the model" versus "true to the data". Covert et al. (2021) show that a large family of methods can be understood as removing features and summarising the effect. Kumar et al. (2020) catalogue problems with Shapley-based importance, including the dependence on the choice of value function and the difficulty of interpreting attributions as actionable. This thesis uses the interventional definition because the purpose of the explanation engine is to describe the model's response to its inputs, and it measures the sensitivity of results to the background distribution.

Gradient-based and other methods. For differentiable models, Integrated Gradients (Sundararajan et al., 2017) attribute predictions by integrating gradients along a path from a baseline and satisfy axioms similar to Shapley's. Global summaries include partial-dependence plots (Friedman, 2001), individual conditional expectation curves (Goldstein et al., 2015) and accumulated local effects (Apley and Zhu, 2020), the last of which handles correlated features more gracefully. Permutation importance (Breiman, 2001) measures the loss increase from shuffling a feature.

2.2.3 Explainability in finance

Bussmann et al. (2021) applied SHAP to credit-risk models for small and medium enterprises and proposed clustering explanations to reveal groups of similar borrowers. Bracke et al. (2019), at the Bank of England, used Shapley values and a related decomposition to explain a mortgage-default model and emphasised the importance of explanation for supervisors. Surveys of deployed XAI (Bhatt et al., 2020) find that in practice explanations are used mainly by engineers for debugging rather than by end users, and that organisations struggle to connect explanation outputs to governance processes. Supervisors have taken note: the Consumer Financial Protection Bureau has stated that creditors using complex algorithms must still provide specific and accurate reasons for adverse action and may not rely on generic checklists when these do not reflect the actual reasons (Consumer Financial Protection Bureau, 2022).

2.3 Evaluating explanations

Evaluation is the weakest link in XAI. Doshi-Velez and Kim (2017) distinguish application-grounded evaluation (real tasks with domain experts), human-grounded evaluation (simplified tasks with lay users) and functionally-grounded evaluation (formal proxies without humans). Functionally-grounded proxies include *deletion* and *insertion* metrics and the retraining-based ROAR benchmark of Hooker et al. (2019), which removes the features ranked most important, retrains the model and measures the drop in accuracy. These

proxies test consistency between an explanation and the model but cannot reveal whether the model's reasons correspond to the true data-generating mechanism.

Robustness is a second axis. Alvarez-Melis and Jaakkola (2018) showed that popular explainers can change dramatically under imperceptible input perturbations and proposed local Lipschitz estimates as a measure of stability. Slack et al. (2020) demonstrated that LIME and SHAP can be fooled: a biased classifier can be wrapped so that its explanations appear innocuous on the perturbed samples the explainers generate. These results imply that an explanation should not be trusted without evidence of stability and fidelity, and that explanation pipelines are themselves an attack surface.

Synthetic data with known ground truth offers a route to fidelity evaluation that real data cannot. It has been used in the attribution literature to validate methods on simple functions, but rarely in a financial setting with realistic structure, interactions, correlated inputs and a known bias mechanism. Chapter 3 develops this approach and introduces a decomposition that separates the error due to the model from the error due to the explainer.

2.4 Counterfactual explanations and algorithmic recourse

Wachter et al. (2017) proposed counterfactual explanations as a way of satisfying the spirit of the GDPR without opening the black box: the explanation states the smallest change to the inputs that would have produced a different outcome. Subsequent work added desiderata. Ustun et al. (2019) introduced *actionable recourse*, restricting changes to features the individual can control and in feasible directions, and showed that some models offer no recourse at all to some individuals. Mothilal et al. (2020) generated diverse sets of counterfactuals (DiCE). Reviews by Verma et al. (2020) and Karimi et al. (2022) list further properties: validity, proximity, sparsity, plausibility (closeness to the data manifold), causality and robustness.

Robustness is especially relevant to autonomous systems, because models are retrained. Pawelczyk et al. (2020) showed that counterfactuals can be invalidated by predictive multiplicity, the existence of many equally accurate models that disagree on individual cases, and Upadhyay et al. (2021) proposed recourse that is robust to model shifts. A recourse recommendation that becomes invalid after the next retraining misleads the customer who follows it. TRACE therefore treats robustness of recourse as a measured property and introduces a validity margin.

2.5 Fairness and proxy discrimination

Machine-learning fairness has produced a range of group criteria. *Demographic parity* requires equal positive-decision rates across groups; *equal opportunity* (Hardt et al., 2016) requires equal true-positive rates, that is, equal approval rates among those who would in fact repay; calibration within groups requires that a given score mean the same risk in every group. Kleinberg et al. (2017) and Chouldechova (2017) proved that these criteria cannot generally be satisfied simultaneously when base rates differ. Barocas and Selbst (2016) explained how disparate impact arises without discriminatory intent, notably through biased historical labels and through proxies: variables correlated with protected characteristics that carry their information into the model even when the protected attribute itself is excluded.

Explanations are a natural tool for detecting proxies, because a proxy that carries bias will receive attribution that is correlated with group membership. This possibility is often mentioned but rarely tested against a known bias mechanism. Chapter 6 tests it.

2.6 Uncertainty quantification and selective decision-making

A model that outputs a probability is only as useful as its calibration. Guo et al. (2017) showed that modern neural networks are often miscalibrated and proposed temperature scaling. Calibration, however, holds on average and does not tell an autonomous system which individual decisions are reliable.

Conformal prediction (Vovk et al., 2005) provides finite-sample, distribution-free guarantees. In split conformal classification, a held-out calibration set is used to compute a threshold on a non-conformity score such that the resulting prediction sets contain the true label with probability at least $1-\alpha$, under exchangeability. Sadinle et al. (2019) characterise the least ambiguous set-valued classifiers with bounded error, and Angelopoulos and Bates (2023) provide an accessible introduction. A key subtlety is that the standard guarantee is *marginal* : it holds averaged over all cases and may fail badly for a minority class. *Mondrian* (class-conditional) conformal prediction (Vovk et al., 2005) calibrates separately within categories and provides a guarantee within each. Prediction sets map naturally onto autonomy: a singleton set signals that the system may act, and an ambiguous or empty set signals that it should defer. This mapping connects conformal prediction to the literature on selective classification and to the levels-of-automation framework.

2.7 Governance, accountability and documentation

Accountability requires more than explanation. Raji et al. (2020) propose an end-to-end internal audit framework spanning the development lifecycle. Documentation artefacts such as model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2021) standardise the information that should accompany a model. In banking, SR 11-7 (Board of Governors of the Federal Reserve System, 2011) has long required effective challenge, independent validation, ongoing monitoring and outcome analysis for models; these requirements map naturally onto explanation and assurance outputs but the mapping is rarely made explicit.

2.8 Regulatory landscape

Four bodies of regulation shape explainable autonomous finance.

Adverse-action rules. Under Regulation B, creditors must disclose the specific principal reasons for adverse action. Consumer Financial Protection Bureau (2022) confirms that this applies to complex algorithms. Reason codes derived from attributions must therefore be accurate for the individual, not merely plausible.

Data protection. The GDPR restricts solely automated decisions with significant effects and provides for human intervention and contestation; the existence and scope of a "right to explanation" has been debated (Goodman and Flaxman, 2017; Wachter et al., 2017).

AI-specific regulation. The EU AI Act (European Parliament and Council, 2024) designates creditworthiness assessment of natural persons as high-risk and requires risk management, data governance, technical documentation, logging, transparency, human oversight and accuracy and robustness.

Model risk management. SR 11-7 (Board of Governors of the Federal Reserve System, 2011) and comparable supervisory expectations require validation, monitoring and governance of models, including those used for investment decisions.

2.9 Synthesis and research gaps

The review supports the three gaps identified in Chapter 1 and refines them.

1. **Ground-truth fidelity in finance.** Explanation fidelity is evaluated against the model or against human judgement, not against the true drivers of financial risk, and the contributions of model error and explainer error are not separated.

2. **Integration.** Attribution, recourse, uncertainty, fairness and governance are studied in largely separate literatures. There is no reference architecture that specifies how they should interact in an autonomous system and how their outputs feed governance.
3. **Robust, actionable recourse in deployed systems.** Actionability and robustness to retraining are recognised desiderata but are rarely measured together in a credit setting.
4. **Guarantees for autonomy.** The mapping from conformal prediction sets to autonomous action is natural but its behaviour under class imbalance, which is endemic in credit, is under-examined.
5. **Decision-level explanation.** Explanations of optimiser-based decisions, as opposed to the predictions that feed them, are largely absent from the literature on financial XAI.

The remainder of the thesis addresses these gaps.

Chapter 3

Research Methodology

3.1 Research paradigm

The research is framed as design science in information systems (Hevner et al., 2004). Design science aims to create and evaluate artefacts that solve identified problems, and it judges them by utility, rigour of construction and rigour of evaluation. The artefact here is TRACE, comprising a conceptual architecture, formal definitions, algorithms and a software instantiation. The process follows the six activities of Peffers et al. (2007): problem identification and motivation (Chapters 1 and 2); definition of objectives (Section 3.2); design and development (Chapter 4); demonstration (Chapters 6 and 7); evaluation (the same chapters and Chapter 8); and communication (this thesis).

3.2 Requirements derivation

Requirements were derived from three sources: the regulatory instruments reviewed in Chapter 2, the desiderata for explanations and recourse in the XAI literature, and the operational needs of autonomous systems (speed, scale, retraining). Table 3.1 lists the resulting requirements and the sources from which each is principally derived.

3.3 Evaluation strategy

3.3.1 Why synthetic data

The core of the evaluation is a set of controlled computational experiments on synthetic data. Real credit data sets would offer realism but have a decisive limitation for this research: the true function relating applicant characteristics to default is unknown. It is therefore impossible to say whether an explanation identifies the true drivers of an applicant's risk, and impossible to separate an inaccurate explanation of a good model from an accurate explanation of a model that has learned the wrong thing. Equally, the true extent of bias in historical labels is unknown in real data, so a proxy-detection method cannot be validated against the right answer.

A synthetic environment overcomes these limitations by construction. The data-generating process, including a non-linear and interacting risk function, correlated fea-

Table 3.1: Requirements for an explainable autonomous financial decision-making system.

ID	Requirement	Principal source
R1	Each autonomous decision must be accompanied by specific, individual reasons that accurately reflect the drivers of that decision.	Reg. B, CFPB, GDPR
R2	The faithfulness of explanations to the model, and of the model to the domain, must be measurable.	SR 11-7, XAI evaluation
R3	Explanations must be stable: small, irrelevant input changes must not substantially change them.	Alvarez-Melis and Jaakkola (2018)
R4	Adversely affected individuals should receive actionable recourse that respects immutable characteristics and feasible directions of change.	Ustun et al. (2019), GDPR
R5	Recourse should remain valid across routine model updates, or its fragility must be disclosed.	Upadhyay et al. (2021)
R6	The system must determine, with stated guarantees, which decisions it may take autonomously and which it must escalate to a human.	AI Act (human oversight)
R7	The system must detect proxy discrimination and report group-level outcomes.	ECOA, AI Act
R8	Domain knowledge (e.g. monotone risk relationships) must be enforceable in the model.	SR 11-7 (conceptual soundness)
R9	Where decisions are produced by an optimiser, explanations must refer to the decision, not only to the underlying predictions.	Gap 5, Chapter 2
R10	Every decision must leave a complete, reproducible audit record.	AI Act (logging), SR 11-7

tures, integer-valued counts, an irrelevant feature, a protected attribute and a proxy, is specified completely (Appendix A). Ground-truth explanations can be computed exactly. Bias can be switched on and off. Experiments can be repeated with different seeds. The approach follows the logic of simulation studies in statistics, where methods are validated in settings where the answer is known before they are trusted where it is not.

3.3.2 Threats arising from the choice

The principal threat is external validity: results on synthetic data may not transfer to real portfolios. The environment was designed to reduce this threat by matching qualitative features of consumer credit data (a default rate around 14%, correlated affordability variables, skewed income, count variables, predictive performance in the range typically reported) and by avoiding a purely additive or linear ground truth that would favour particular methods. Chapter 8 discusses the remaining threats in detail. The thesis does not claim that the specific numerical results will hold on any particular real portfolio; it claims that the phenomena demonstrated (for example, that model error dominates explainer error, or that marginal conformal guarantees can hide class-level failure) are

mechanisms that practitioners should test for.

3.3.3 Evaluation criteria

Each requirement in Table 3.1 is linked to one or more measurable criteria, defined formally in Chapters 4 and 5:

predictive performance and calibration (AUC, Brier score, expected calibration error) for R8;

fidelity to ground truth and to the model (rank correlation of absolute attributions, top-3 agreement, sign agreement, relative ℓ_2 error) for R1 and R2;

stability (agreement across random seeds; relative local Lipschitz constant) for R3;

recourse validity, cost, sparsity, actionability, plausibility and post-retraining validity for R4 and R5;

conformal coverage, overall and within classes and groups, escalation rate and accuracy of autonomous decisions for R6;

group approval-rate gaps, equal-opportunity gaps, proxy importance and attribution–group correlation for R7;

decision-level attribution efficiency and the comparison of forecast-level with decision-level importance for R9;

completeness of the audit record against the schema for R10.

3.4 Ground-truth attribution

Let $f^\star : \mathrm{X} \to \mathrm{R}$ denote the true log-odds of default as a function of the observed features, and let $\hat{f}$ denote the log-odds of a fitted model. For an instance $x \in \mathrm{R}^d$ and a background sample $B = \{b^{(1)}, \ldots, b^{(m)}\}$ drawn from the training distribution, the interventional coalition value of a feature subset $S \subseteq \{1, \ldots, d\}$ is

$$v_{g,x}(S) = \frac{1}{m} \sum_{k=1}^{m} g\left(x_S, b_S^{(k)}\right), \tag{3.1}$$

where (x_S, b_{S^c}) denotes the vector taking the values of x on S and of b elsewhere, and $g \in \{f^\star, \hat{f}\}$. The Shapley value of feature j is

$$\phi_j(g, x) = \sum_{S \subseteq \{1..d\} \setminus \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \left[v_{g,x}(S \cup \{j\}) - v_{g,x}(S) \right]. \tag{3.2}$$

With $d = 11$ features there are $2^{11} = 2048$ coalitions, so (3.2) can be computed exactly by enumeration. The *ground-truth attribution* is $\phi(f^{\star}, x)$; the *model's exact attribution* is $\phi(\hat{f}, x)$; an explainer E produces an estimate $\phi^{E}(\hat{f}, x)$.

Definition 3.1 (Model gap and explainer gap)**.** *For a dissimilarity D between attribution vectors, the total explanation error* $D(\phi^{E}(\hat{f}, x), \phi(f^{\star}, x))$ *is bounded, for metric D, by the sum of the* model gap $D(\phi(\hat{f}, x), \phi(f^{\star}, x))$ *and the* explainer gap $D(\phi^{E}(\hat{f}, x), \phi(\hat{f}, x))$.

The decomposition matters because the remedies differ. A large explainer gap calls for a better explainer or a larger sampling budget. A large model gap calls for a better or more constrained model, more data, or a change of model class, and it means that even a perfect explanation will give customers reasons that do not reflect the true drivers of their risk.

3.5 Software and reproducibility

All experiments were implemented in Python using scikit-learn (Pedregosa et al., 2011), the shap package for KernelSHAP and TreeSHAP, the lime package, SciPy for optimisation, and custom code for exact Shapley enumeration, counterfactual search, conformal prediction, decision-level attribution and audit records. All random processes are seeded, and the seeds are listed in Appendix B. Experiments were run on a single CPU core, and run times are reported where relevant to indicate operational cost.

3.6 Ethical considerations

Because no real personal data were used and no human participants were involved, the research raised no data-protection or participant-welfare issues. The synthetic protected attribute is an abstract binary group label and is not intended to model any real demographic group. The research nevertheless engages with ethical questions about automated decisions, and these are discussed in Chapter 8.

Chapter 4

The TRACE Framework

This chapter presents TRACE, the framework developed in this thesis. The name stands for the five properties the framework is designed to secure: decisions that are **T**ransparent, **R**ecourse-aware, **A**ccountable, **C**alibrated and **E**xplainable. The chapter first gives an overview of the architecture and then describes each layer, giving formal definitions and algorithms where needed.

4.1 Design principles

Five principles guided the design.

1. **Explain the decision, not just the score.** Explanations are computed for the action the system takes. Where the action is a threshold on a score the two coincide up to the threshold; where it is the output of an optimiser they do not (requirement R9).
2. **Assure before you explain.** An explanation is released only with evidence about its fidelity and stability. Explanations are treated as claims to be verified, not as self-evidently true (R2, R3).
3. **Constrain by design.** Domain knowledge is encoded in the model through constraints such as monotonicity, which reduces the need for post-hoc correction and makes explanations more coherent (R8).
4. **Autonomy is earned per decision.** The system acts autonomously only when a calibrated uncertainty test says it may; otherwise it escalates (R6).
5. **Everything is recorded.** Inputs, model version, scores, explanations, recourse, uncertainty and the routing decision are written to an immutable audit record (R10).

4.2 Architecture overview

Figure 4.1 shows the four layers and the flow of information among them. A request (a credit application, or a scheduled rebalancing event) enters the *decision engine*, which

produces a score and a candidate action. The *explanation engine* computes local attributions, decision-level attributions and, for adverse outcomes, recourse. The *assurance layer* computes conformal prediction sets and decides whether the case may be decided autonomously, checks explanation fidelity and stability on a monitored sample, and runs periodic fairness and proxy audits. The *governance layer* converts these outputs into customer-facing reason codes and recourse statements, writes the audit record, routes escalated cases to human reviewers and feeds monitoring dashboards and model-risk reports. Human reviewers' decisions return to the system as labelled data and as challenges to the model.

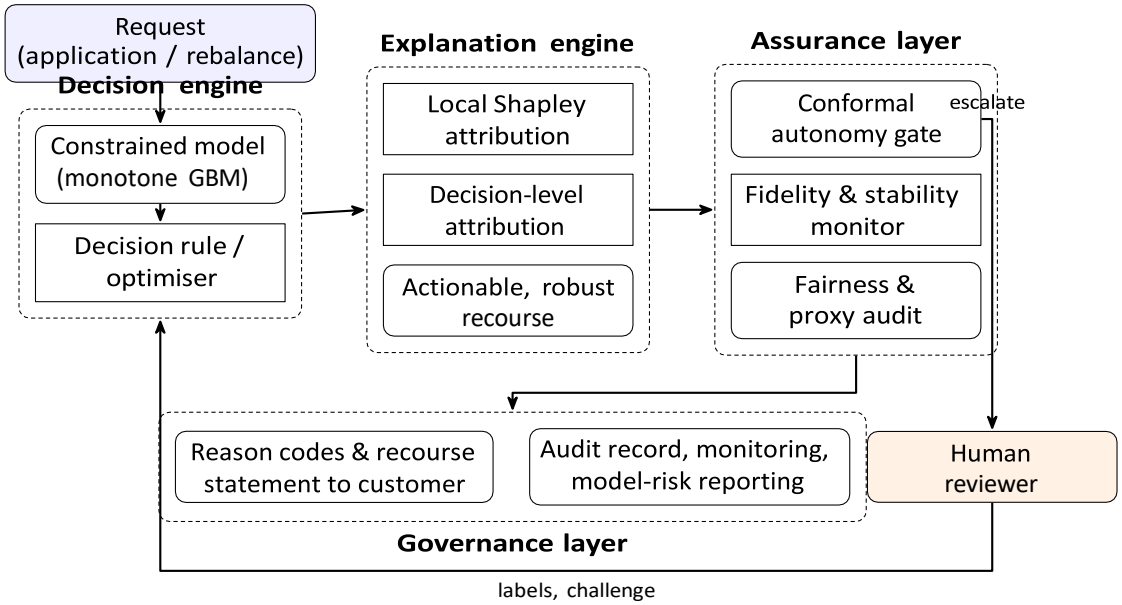

Figure 4.1: The TRACE architecture: four layers and the principal information flows.

4.3 Layer 1: Decision engine

4.3.1 Constrained scoring models

The decision engine contains the predictive model and the rule or optimiser that converts predictions into actions. TRACE does not mandate a model class, but it requires that the model support the explanation and assurance layers and that domain constraints be enforceable. The default instantiation is a histogram-based gradient-boosting classifier with *monotonic constraints*. For each feature j, a sign $c_j \in \{-1, 0, +1\}$ is specified, and the model is constrained so that, holding other features fixed, its output is non-decreasing in x_j if $c_j = +1$ and non-increasing if $c_j = -1$. In credit, for example, risk should not fall as the debt-to-income ratio rises, nor rise as savings increase.

Monotonicity serves several purposes at once. It encodes domain knowledge, which supports conceptual soundness under model-risk rules. It prevents explanations that

would be incoherent to a customer, such as "your application was declined partly because your income is too high". And it simplifies recourse: if risk is monotone in an actionable feature, moving that feature in the favourable direction can never hurt.

4.3.2 Decision rules

For a binary decision such as credit approval, the rule is a threshold: approve if $p\hat{}(x) < \tau$. The threshold reflects risk appetite and profitability and is a governance parameter. For allocation decisions, the decision is the solution of an optimisation problem whose inputs are model predictions. In the portfolio instantiation, expected returns $\hat{\mu}$ come from a forecasting model and the weights solve

$$w^{\star}(\hat{\mu}, \Sigma) = \arg\max_{w} \; \hat{\mu}^{\top} w - \frac{\gamma}{2} w^{\top} \Sigma w \qquad \text{s.t.} \quad \mathbf{1}^{\top} w = 1,\; 0 \le w_{i} \le \bar{w}, \tag{4.1}$$

with risk aversion γ, covariance estimate Σ and position cap $\bar{w}$.

4.4 Layer 2: Explanation engine

4.4.1 Local attribution

For each decision the engine computes interventional Shapley values (3.2) of the model's log-odds with respect to a fixed, versioned background sample. Log-odds rather than probabilities are explained because they are additive in the attributions of a logistic model and because contributions in log-odds space do not depend on the level of the other contributions. For tree ensembles, attributions are computed exactly with interventional TreeSHAP; for other models, KernelSHAP is used with a sampling budget chosen by the fidelity monitor (Section 4.5.1); and for low-dimensional problems exact enumeration is feasible.

By efficiency, the attributions satisfy $\sum_{j} \phi_{j}(\hat{f}, x) = \hat{f}(x) - \mathrm{E}_{B}[\hat{f}]$, so they decompose the difference between the applicant's score and the average score of the background population. This is the quantity a customer implicitly asks about: "why is my risk higher than usual?"

4.4.2 Reason codes

Adverse-action notices require the principal reasons for the decision. TRACE derives them from the local attribution as follows. Features with positive attribution (pushing towards decline) are ranked by magnitude; the top k (typically four) are mapped through a maintained dictionary to customer-facing statements, each qualified by the direction of the applicant's value relative to the reference population. Two safeguards apply.

Features whose attribution is below a materiality threshold are not reported, so that a customer is not given an immaterial reason; and features designated as non-reportable (for example, a proxy flagged by the fairness audit) trigger an escalation rather than a reason code.

4.4.3 Actionable and robust recourse

For an applicant who is declined, the recourse problem is

$$\min_{x'} \; c(x, x') = \sum_{j} \frac{|x'_j - x_j|}{\mathrm{MAD}_j} \quad \text{s.t.} \quad \hat{p}(x') < \tau - \delta, \; x'_j = x_j \;\; \forall j \in \mathsf{I}, \;\; x'_j \in [\ell_j(x), u_j(x)] \;\; \forall j \in \mathsf{A}, \tag{4.2}$$

where I is the set of immutable features (such as age, past late payments and credit history length), A is the set of actionable features, each with a feasible interval $[\ell_j(x), u_j(x)]$ that encodes the permitted direction and magnitude of change (for example, income may rise by at most 20%, utilisation may only fall), MAD_j is the median absolute deviation of feature j in the training data, which puts costs on a common scale, and $\delta \geq 0$ is a *validity margin*. The cost is an ℓ_1 norm, which favours sparse changes.

Because tree ensembles are piecewise constant, (4.2) is non-differentiable. TRACE solves it with a sampling-and-shrinking heuristic (Algorithm 1) that is model-agnostic, respects integer features, and returns sparse solutions. The validity margin δ addresses requirement R5: by requiring the counterfactual to clear the threshold by a margin, it increases the chance that the recourse remains valid after the model is retrained, at the price of a higher cost.

4.4.4 Decision-level attribution

When the action is the output of an optimiser, an explanation of each forecast does not explain the action. TRACE therefore defines attribution directly on the decision. Let the inputs be partitioned into G interpretable *groups* $\mathsf{G} = \{g_1, \ldots, g_G\}$ (for example valuation, momentum, risk and macroeconomic signals), let x be the current input matrix (one row per asset) and let $\bar{x}$ be a reference input (the training mean). Define the decision under a coalition of groups $T \subseteq \mathsf{G}$ as

$$w(T) = w^{\star}\big(\hat{\mu}(x_T, \bar{x}_T), \Sigma\big), \tag{4.3}$$

the weight vector the optimiser would choose if only the signals in T took their current values. The decision-level attribution of group g to the weight of asset i is the Shapley value

$$\psi_{g,i} = \sum_{T \subseteq \mathsf{G} \setminus \{g\}} \frac{|T|!(G - |T| - 1)!}{G!} \Big(w_i(T \cup \{g\}) - w_i(T) \Big). \tag{4.4}$$

Algorithm 1: TRACE-CF: actionable counterfactual search

Input: instance x, model $\hat{p}$, threshold τ, margin δ, actionable set $\mathcal{A}$ with bounds, budget n

Output: counterfactual x' or NONE

for $i = 1$ **to** n **do**
  draw $k \sim \text{Uniform}\{1, 2, 3\}$ and a subset $J \subseteq \mathcal{A}$ with $|J| = k$;
  $c^{(i)} \leftarrow x$; **foreach** $j \in J$ **do** $c_j^{(i)} \sim \text{Uniform}[\ell_j(x), u_j(x)]$;
  round integer features of $c^{(i)}$;
$V \leftarrow \{c^{(i)} : \hat{p}(c^{(i)}) < \tau - \delta\}$;
if $V = \varnothing$ **then return** NONE;
foreach *of the 20 lowest-cost candidates* $c \in V$ **do**
  repeat
    foreach *feature* j *with* $c_j \neq x_j$ **do**
      find the smallest $\alpha \in \{0.1, 0.2, \ldots, 1\}$ with $\hat{p}(x + \alpha\, e_j(c_j - x_j) \text{ on } j) < \tau - \delta$ and set $c_j \leftarrow x_j + \alpha(c_j - x_j)$;
  until *three passes completed*;
return the shrunk candidate with minimal cost $c(x, \cdot)$;

Proposition 4.1 (Efficiency of decision-level attribution)**.** *For each asset i, $\sum_{g \in G} \psi_{g,i} = w_i(\mathcal{G}) - w_i(\varnothing)$: the group attributions sum exactly to the difference between the actual weight and the reference weight.*

Proof. This is the efficiency property of the Shapley value applied to the game $T \mapsto w_i(T)$. □

Because G is small, the 2^G optimisations needed are cheap. The approach handles the non-linearity of the optimiser, including binding constraints, without approximation, and it allows the question "why did the system move into this asset?" to be answered in terms of signal families that portfolio managers and supervisors understand.

4.5 Layer 3: Assurance layer

4.5.1 Fidelity and stability monitoring

The assurance layer treats each explanation method as a component with measurable error. On a monitoring sample of decisions, it computes exact Shapley values (by enumeration or TreeSHAP) and compares the production explainer with them, reporting the rank correlation of absolute attributions, top-3 agreement and sign agreement. If agreement falls below a tolerance, the sampling budget is raised or the explainer is replaced. In a simulation setting, or wherever a trusted reference model is available, the same metrics are computed against the reference to estimate the model gap.

Stability is monitored in two ways: by recomputing explanations with different random seeds, and by estimating a relative local Lipschitz constant

$$L(x) = \max_{x' \in N(x)} \frac{\|\phi(x') - \phi(x)\|_2 / \|\phi(x)\|_2}{\|(x' - x)/s\|_2}, \tag{4.5}$$

where $N(x)$ is a set of small perturbations of x and s is the vector of feature standard deviations. Large values indicate that explanations change disproportionately with immaterial changes in the input.

4.5.2 Conformal autonomy gate

The gate decides whether a case may be decided autonomously. Let $\hat{\pi}_y(x)$ be the model's probability for class $y \in \{0, 1\}$ and define the non-conformity score $s(x, y) = 1 - \hat{\pi}_y(x)$. On a calibration set $\{(x_i, y_i)\}_{i=1}^{n}$ disjoint from training, compute the quantile

$$\hat{q} = \text{the } \lceil (n+1)(1-\alpha) \rceil \text{-th smallest of } s(x_1, y_1), \ldots, s(x_n, y_n). \tag{4.6}$$

The prediction set for a new case is $C(x) = \{y : s(x, y) \leq \hat{q}\}$. Under exchangeability, $\Pr\{y \in C(x)\} \geq 1 - \alpha$ (Vovk et al., 2005). In the *Mondrian* variant, the quantile $\hat{q}_y$ is computed separately from the calibration cases with label y, giving $C(x) = \{y : s(x, y) \leq \hat{q}_y\}$ and the stronger guarantee $\Pr\{y \in C(x) \mid Y = y\} \geq 1 - \alpha$ for each class.

The gate's rule is simple. If $|C(x)| = 1$, the case is decided autonomously according to the single label in the set (combined with the business threshold where these coincide). If $|C(x)| = 2$ (both outcomes plausible) or $|C(x)| = 0$ (neither outcome conforms, signalling an atypical case), the case is escalated to a human reviewer together with its explanation. The choice between marginal and Mondrian calibration is a governance decision: Chapter 6 shows that under class imbalance it has large consequences.

4.5.3 Fairness and proxy audit

Periodically, and before any model is deployed, the assurance layer computes group approval rates, the demographic-parity gap $\Delta_{DP} = \Pr(\text{approve} \mid A = 0) - \Pr(\text{approve} \mid A = 1)$ and the equal-opportunity gap Δ_{EO}, the difference in approval rates among applicants who are truly low-risk. Where the protected attribute A is available for auditing (even though it is not used by the model), the layer also computes, for each feature j, the mean absolute attribution and the correlation $\rho_j = \text{corr}(\phi_j(\hat{f}, x), A)$ across applicants. A feature with both material importance and a high ρ_j is flagged as a potential proxy and referred for review. The rationale is that if a feature carries information about A into the model's decisions, its contributions will differ systematically between groups.

4.6 Layer 4: Governance layer

The governance layer turns the outputs of the other layers into organisational artefacts.

Customer communications. Reason codes and, where available, a recourse statement ("your application would have been approved if your credit utilisation were below 31% and your income were at least 29,100") are produced for every adverse decision.

Audit record. For every decision an immutable record is written containing a request identifier, timestamp, model and background-sample versions, inputs, score, threshold, prediction set and gate outcome, attributions, reason codes, recourse and its margin, and the final action and actor (system or named reviewer). The schema is given in Appendix C.

Escalation workflow. Escalated cases are queued for human review with the explanation attached; reviewers' decisions and rationales are recorded and fed back as training data and as evidence for model challenge.

Monitoring and reporting. Fidelity, stability, coverage, escalation rates, group outcomes and attribution drift are tracked over time and reported to model-risk management.

Table 4.1 maps the framework components to the requirements and regulatory obligations they address.

Table 4.1: Mapping of TRACE components to requirements and obligations.

Component	Obligation addressed	Req.
Monotone constrained model	Conceptual soundness (SR 11-7); coherent reasons	R8
Local Shapley + reason codes	Specific principal reasons for adverse action (Reg. B, CFPB)	R1
Fidelity monitor	Validation of explanation method; accuracy of reasons	R2
Stability monitor	Robustness (AI Act); reliability of reasons	R3
TRACE-CF recourse	Meaningful information and contestability (GDPR)	R4
Validity margin	Reliability of recourse across model updates	R5
Conformal autonomy gate	Human oversight (AI Act Art. 14); levels of automation	R6
Fairness & proxy audit	Non-discrimination (ECOA); data governance (AI Act)	R7
Decision-level attribution	Explaining optimiser-based investment decisions	R9
Audit record	Record-keeping and logging (AI Act); model inventory	R10

4.7 Summary

TRACE answers RQ1 by specifying a layered architecture in which explanation is one of several interacting mechanisms rather than an afterthought. The next chapter describes

the environment in which the framework was instantiated and tested.

Chapter 5

Experimental Environment and Protocol

This chapter describes the synthetic environments, the models, the explanation methods and the protocol for each experiment. Full specifications are given in Appendices A and B.

5.1 Synthetic credit environment

5.1.1 Features

Each applicant is described by eleven features available to the model and one protected attribute that is not. The protected attribute $A \in \{0, 1\}$ is drawn with $\Pr(A = 1) = 0.35$. The features are: age (uniform on 21–75 years); income (log-normal, in thousands, with lower mean for group $A = 1$ and a mild age gradient); dti, the debt-to-income ratio (a scaled beta variable, slightly higher on average for $A = 1$); utilisation of revolving credit (beta); late_payments (Poisson, with rate increasing in utilisation); history_years, the length of credit history (increasing in age); open_accounts (Poisson); loan_amount (proportional to income times a uniform factor); inquiries, the number of recent credit inquiries (Poisson); savings_ratio (beta); and postcode_index, a continuous index of residential area constructed as $0.8A$ plus Gaussian noise. The postcode index thus has a correlation of about 0.6 with the protected attribute but no causal role in default.

5.1.2 True risk function

The true log-odds of default are

$$\begin{aligned} f^*(x) = 1.4 \quad & -2.0 + 3.0(\text{dti} - 0.30) + 2.0(\text{util} - 0.40) + 0.55\,\text{late} - 0.9\ln(\text{income}/45) \\ & - 0.06(\text{history} - 10) + 0.25(\text{inq} - 1) - 3.0(\text{savings} - 0.25) \\ & + 1.2(\text{loan}/\text{income} - 0.35) + 4.0(\text{dti} - 0.30)(\text{util} - 0.40)\,, \end{aligned} \tag{5.1}$$

and default is drawn as $Y \sim \text{Bernoulli}(\sigma(f^\star(x) + \beta A))$ with $\beta = 0$ in the clean environment. The function has several deliberate properties. It is non-linear in income (logarithmic). It contains an interaction between dti and utilisation. It depends on the *ratio* of loan amount to income, so neither of these features has an additive effect in the raw feature space. And three of the eleven observed features (age, open_accounts and postcode_index) have no direct effect, although age is correlated with history_years. The resulting default rate is 14.3%.

5.1.3 Biased environment

To study proxy discrimination, a second data set is drawn with $\beta = 0.8$: members of group $A = 1$ have their recorded default outcomes inflated relative to their true creditworthiness. This mimics historical label bias, for instance from past lending practices that set members of a group up to fail, or from biased recording of outcomes. The model never sees A, but it can learn the bias through the postcode proxy.

5.1.4 Data splits

The clean environment comprises 40,000 applicants: 24,000 for training, 6,000 for conformal calibration and 10,000 for testing. The biased environment comprises 40,000 applicants, with 30,000 for training and 10,000 for testing; a clean environment with the same seed provides the reference.

5.2 Models

Four models were trained on the clean data.

Logistic regression on standardised features, the traditional scorecard benchmark.

Gradient boosting (GBM): 200 trees of depth 3, learning rate 0.05.

Monotone HGB: a histogram-based gradient-boosting classifier with 300 iterations, at most 15 leaves per tree, learning rate 0.05, and monotonic constraints: increasing in dti, utilisation, late_payments, loan_amount and inquiries; decreasing in income, history_years and savings_ratio; unconstrained in age, open_accounts and postcode_index. This is the default TRACE decision model.

MLP: a neural network with hidden layers of 64 and 32 units, ℓ_2 penalty 10^{-3} and early stopping.

The true model $\sigma(f^\star)$ serves as an upper benchmark. The decision threshold is $\tau = 0.25$, which declines 17.4% of test applicants under the monotone HGB.

5.3 Explanation methods

All explanations are computed on the log-odds scale, except LIME, which explains probabilities; for LIME, comparisons are made on rank-based metrics only.

Exact Shapley by enumeration of all 2048 coalitions using (3.1)–(3.2), with a background of 50 training applicants.

TreeSHAP (interventional) for the GBM, with the same background.

KernelSHAP with sampling budgets of 200 and 40 coalitions, with the same background.

LIME with 1000 perturbation samples, continuous features not discretised, and the coefficient for feature j converted to a contribution by multiplying by the standardised deviation of x_j from the training mean.

Ground-truth attributions are exact Shapley values of $f^\star$ with the same background.

5.4 Metrics

For attribution vectors a (explanation) and b (reference) of one instance:

Rank correlation: Spearman correlation between $|a|$ and $|b|$, which measures whether the explanation ranks features by importance correctly;

Top-3 agreement: $|\text{top}_3(|a|) \cap \text{top}_3(|b|)|/3$, which corresponds to whether the principal reasons reported to a customer are correct;

Sign agreement: the share of features with $|b_j| > 0.05$ for which $\text{sign}(a_j) = \text{sign}(b_j)$, which corresponds to whether a reason is reported in the right direction;

Relative ℓ_2 error: $\|a - b\|_2 / \|b\|_2$.

Metrics are averaged over 200 randomly selected test applicants. Ninety-five per cent confidence intervals for the mean rank correlation are obtained by bootstrapping over applicants with 1000 resamples.

Predictive performance is measured by AUC, Brier score and expected calibration error (ECE) with ten equal-width bins. Recourse is measured by validity (share of declined applicants for whom a counterfactual was found), cost in MAD units, sparsity (number of features changed), the share of counterfactuals that are *non-actionable* (requiring a change to an immutable feature or a change in an infeasible direction), plausibility (distance in standardised space to the nearest training applicant, divided by the typical nearest-neighbour distance, so that 1 means as close to the data as a typical real applicant), and robust validity (the share of counterfactuals that remain approved by a model retrained on a bootstrap resample with a different seed).

5.5 Experiment protocols

E1: Predictive performance. The four models and the true model are evaluated on the test set.

E2: Explanation fidelity. For 200 test applicants, ground-truth attributions and each model's exact, KernelSHAP and LIME attributions (and TreeSHAP for the GBM) are computed. Each explanation is compared with the ground truth (total error) and with the model's exact attribution (explainer gap). Sensitivity to the background sample is assessed by recomputing GBM exact attributions for 60 applicants with five independently drawn backgrounds.

E3: Explanation stability. For 40 applicants, LIME and KernelSHAP (40 samples) are run with four random seeds and pairwise agreement is computed. For 25 applicants, five small perturbations (Gaussian noise with standard deviation 2% of the feature standard deviation on the continuous features) are applied and the relative local Lipschitz constant is computed for exact Shapley and for LIME, using the monotone HGB.

E4: Recourse. For the first 150 declined test applicants, three methods are compared: an unconstrained nearest counterfactual that may change any feature in either direction (a Wachter-style baseline implemented with the same search procedure), TRACE-CF with actionability constraints, and TRACE-CF with a validity margin $\delta = 0.05$. Actionable features and bounds are: income may rise by up to 20%; dti and utilisation may only fall; loan amount may fall to 40% of the request; savings may rise up to 0.95. All other features are immutable.

E5: Conformal autonomy gate. Marginal and Mondrian split conformal prediction are calibrated on the 6000 calibration applicants and evaluated on the test set for $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$, reporting coverage overall, by class and by group, escalation rate and the accuracy of autonomous decisions.

E6: Fairness and proxy audit. Monotone HGB models are trained on (i) clean labels, (ii) biased labels with all features, and (iii) biased labels with the postcode index removed. Approval rates by group, the demographic-parity gap, the equal-opportunity gap among truly low-risk applicants (true default probability below 15%), and the AUC against outcomes drawn from the clean risk function are computed. KernelSHAP attributions for 300 test applicants yield feature importance and attribution–group correlations.

E7: Autonomous portfolio allocation. Described in Section 5.6.

5.6 Synthetic market environment

Eight assets are simulated over 1100 weeks. A market factor has weekly mean 0.12% and volatility 2%, and its mean is reduced in proportion to a persistent credit-spread state variable (an AR(1) process with coefficient 0.97). Each asset has a market beta between 0.6 and 1.4. Excess returns contain two genuine but weak predictable components: a persistent latent valuation signal (AR(1), coefficient 0.95) with loading 0.25% per unit, and a momentum component $0.15\% \times \tanh(10\, m_{26})$, where m_{26} is the asset's trailing 26-week return. Idiosyncratic noise has weekly volatility 2.5%. The valuation signal is observed only with noise of the same variance as its innovations.

The forecaster sees six features per asset: the noisy valuation signal, 4-week and 26-week momentum, 13-week volatility, the credit spread (common to all assets) and a 52-week rolling beta. A gradient-boosting regressor (150 trees of depth 3, learning rate 0.05, subsampling 0.8) is trained on all data available up to each retraining date, starting at week 312 and retraining every 52 weeks; the out-of-sample period covers 787 weeks. Each week, the agent solves (4.1) with $\gamma = 5$, a cap of 30% per asset, and a covariance matrix estimated from the trailing 104 weeks and shrunk 30% towards its diagonal. Benchmarks are equal weights, inverse-volatility weights and mean–variance weights using the trailing 104-week historical mean. Performance is measured net of transaction costs of 10 basis points per unit of turnover.

Explanations are computed at two levels. At the forecast level, interventional Tree-SHAP attributions of the forecaster over the preceding 52 weeks are computed at each retraining date. At the decision level, the six features are grouped into four signal families (valuation; momentum = 4- and 26-week momentum; risk = volatility and beta; macro = credit spread), and the decision-level attribution of Chapter 4 is computed at every weekly decision with the training mean as reference.

Chapter 6

Results I: Autonomous Credit Decisions

This chapter reports experiments E1–E6. Each section states the result, interprets it and links it to the research questions.

6.1 E1: Predictive performance of constrained and flexible models

Table 6.1 reports test-set performance. All four models come close to the true model, whose AUC of 0.850 is the ceiling set by the irreducible randomness of default. The spread among the fitted models is only 0.005 AUC. Logistic regression performs strongly, because the true function, while non-linear, is dominated by smooth monotone effects; it is also the best calibrated. The monotone HGB is marginally better than the unconstrained GBM on every metric and trains twenty times faster. The MLP achieves the highest AUC of the fitted models, but by a margin that is negligible in practice.

Table 6.1: E1: Test-set predictive performance (n = 10,000).

Model	AUC	Brier	ECE	Fit time (s)
Bayes (true model)	0.850	0.0886	0.0074	0.00
Logistic regression	0.847	0.0891	0.0044	0.02
Gradient boosting	0.843	0.0908	0.0128	9.55
Monotone HGB	0.845	0.0900	0.0090	0.45
MLP	0.848	0.0895	0.0111	0.93

The result supports design principle 3 (constrain by design). Imposing monotonicity, which encodes domain knowledge and guarantees coherent reasons, costs nothing measurable here, and indeed slightly improves performance relative to the unconstrained ensemble, consistent with the view that constraints act as a useful regulariser. It also echoes Rudin (2019): in this environment, the flexible models offer no meaningful accuracy advantage over more interpretable alternatives.

6.2 E2: Explanation fidelity against ground truth

6.2.1 Total fidelity

Table 6.2 and Figure 6.1 report agreement between each explanation and the ground-truth attribution of the true risk function. Three patterns stand out.

First, fidelity varies more across models than across explainers. The best achievable explanation of each model, its exact Shapley attribution, agrees with the ground truth at a rank correlation of 0.960 for logistic regression, 0.921 for the MLP, 0.904 for the monotone HGB and only 0.870 for the unconstrained GBM. In terms of the principal reasons a customer would receive, the top three features are correct 92.5% of the time for logistic regression but only 80.2% of the time for the GBM. Put plainly, for about one reason in five, a perfectly computed explanation of the GBM gives the customer a reason that is not among the three true principal drivers of their risk.

Second, KernelSHAP with 200 coalitions and TreeSHAP reproduce the model's exact attribution almost perfectly, so their total fidelity is indistinguishable from that of exact Shapley. With only 40 coalitions, KernelSHAP loses some fidelity, most for the tree models.

Third, LIME is consistently the least faithful: its rank correlation with the ground truth ranges from 0.771 (GBM) to 0.862 (logistic regression), and its sign agreement is lower, which means it more often reports a reason in the wrong direction.

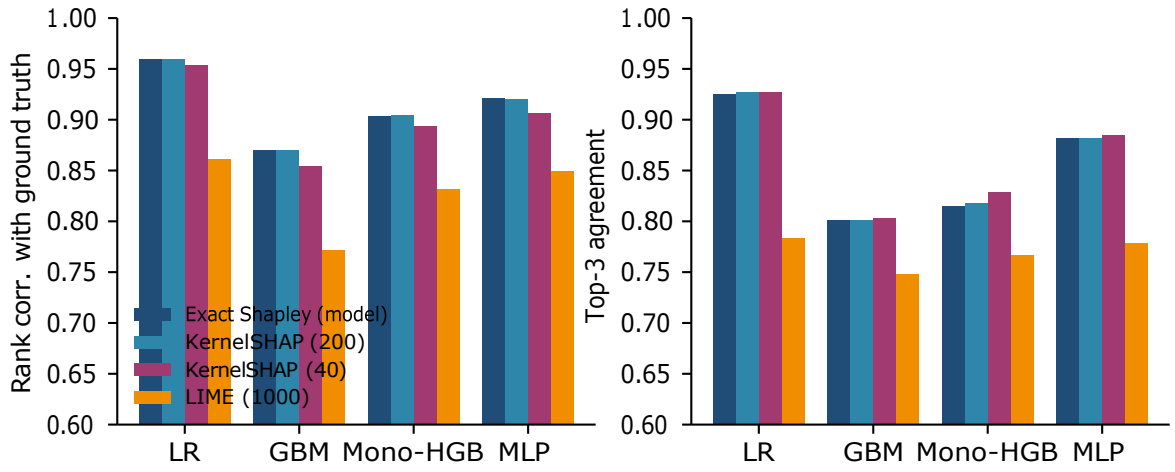

Figure 6.1: E2: Agreement of explanations with the ground-truth attribution, by model and explainer.

6.2.2 Decomposing the error: model gap versus explainer gap

Table 6.3 and Figure 6.2 separate the two sources of error. The model gap (one minus the rank correlation of the model's exact attribution with the truth) ranges from 0.040

to 0.130. The explainer gap of KernelSHAP-200 is an order of magnitude smaller: its rank correlation with the model's exact attribution is at least 0.994 for every model. TreeSHAP is exact. KernelSHAP-40 has an explainer gap of 0.014–0.036, comparable to the model gap of the best models. LIME's explainer gap (0.123–0.179) is larger than the model gap of every model.

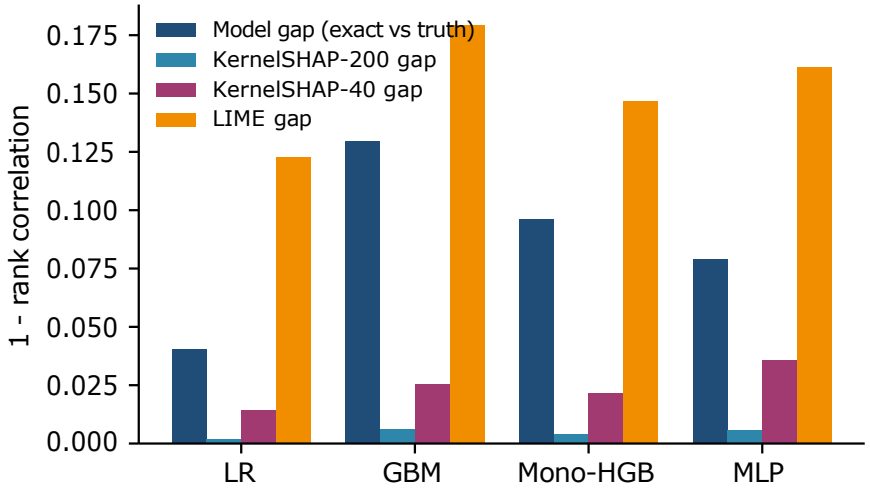

Figure 6.2: E2: Model gap compared with the explainer gaps of KernelSHAP and LIME.

This decomposition answers RQ2. With a modern Shapley-based explainer and an adequate budget, *the dominant source of explanation error is the model, not the explainer*. This has a practical implication that is easy to miss. Institutions often validate explanation methods by checking that they are consistent with the model, for instance by comparing KernelSHAP with TreeSHAP. Such checks would pass here with near-perfect scores even for the GBM, whose explanations nevertheless misidentify the true principal drivers for a material share of customers. Faithfulness to the model is necessary but not sufficient for faithful reasons.

Why does the GBM have the largest model gap despite a competitive AUC? Figure 6.3 compares mean absolute attributions. The GBM distributes attribution in ways that differ from the truth. It attaches weight to age (mean absolute attribution 0.062 against a true value of zero), which is correlated with credit history but has no direct effect, while under-attributing credit history itself (0.326 against 0.537). Most strikingly, it captures the loan-to-income ratio poorly: loan_amount receives a mean absolute attribution of 0.080 against a true 0.370, and income 0.407 against 0.752. The monotone HGB is closer to the truth on every one of these features (for example 0.546 for income and 0.139 for loan amount), although it shares the spurious attribution to age. Many functions achieve similar AUC (the "Rashomon effect"), and they need not agree on the reasons. The monotone constraints narrow the set of admissible functions towards ones consistent with domain knowledge, which reduces the model gap by about a quarter

(from 0.130 to 0.096) at no cost in accuracy.

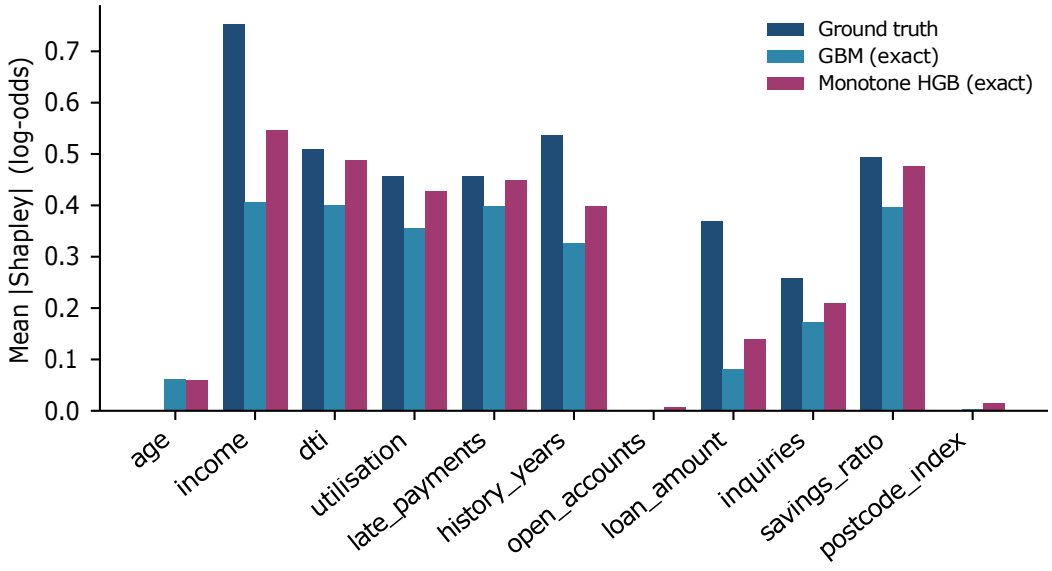

Figure 6.3: E2: Mean absolute attribution by feature for the ground truth, the GBM and the monotone HGB.

6.2.3 Sensitivity to the background sample

Interventional Shapley values depend on the background sample. Recomputing the GBM's exact attributions for 60 applicants with five independent backgrounds of 50 applicants gave a mean standard deviation across backgrounds of 0.035 log-odds, against a mean absolute attribution of 0.248, that is about 14% relative variability. The rank ordering was stable (mean rank correlation across backgrounds 0.951). The background is therefore a material modelling choice that should be versioned and held fixed in production, as TRACE requires, and backgrounds of more than 50 observations are advisable where reason codes are close to the materiality threshold.

6.3 E3: Explanation stability

Table 6.4 reports stability for the monotone HGB. Across random seeds, LIME and KernelSHAP-40 have similar rank-correlation stability (0.972 and 0.970), but KernelSHAP is more stable in its top three features (0.976 against 0.947): with LIME, roughly one in twenty principal reasons changes simply because the explainer is rerun. Exact Shapley is deterministic by construction.

Under small input perturbations, the median relative local Lipschitz constant of LIME (3.55) is more than twice that of exact Shapley (1.59). An immaterial change to an applicant's data thus changes LIME's explanation proportionally more than twice as

much. Because the model itself is piecewise constant, some sensitivity is inherent; the excess sensitivity of LIME arises from its sampling and surrogate fitting.

Together, E2 and E3 support the design choice of exact or near-exact Shapley attribution in the TRACE explanation engine, and they imply that LIME, at least in its default configuration, is not suitable as the basis of legally required reason codes.

6.4 E4: Actionable and robust recourse

Table 6.5 and Figure 6.4 compare recourse methods for 150 declined applicants. All three methods found a counterfactual for every applicant, so validity is not the differentiator.

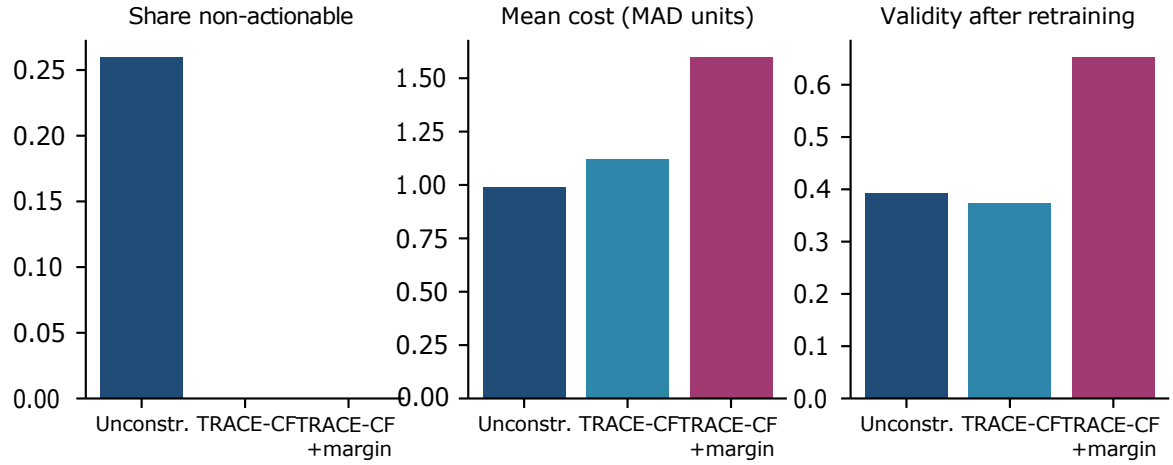

Figure 6.4: E4: Actionability, cost and robustness of recourse.

Actionability. The unconstrained counterfactual is the cheapest (0.99 MAD units), but 26% of its recommendations are non-actionable: they require the applicant to change an immutable characteristic, such as age or history of late payments, or to move a feature in an infeasible direction. Such recourse is worse than none, because it is either impossible to follow or, in the case of age, discriminatory in effect. TRACE-CF produces no non-actionable recommendations, at a 13% higher cost and with slightly more features changed (1.95 against 1.55). Plausibility is essentially identical across methods (close to 1.0), meaning the recommended profiles are as close to real applicants as real applicants are to each other.

Robustness. The most important finding concerns robustness. When the model is retrained on a bootstrap resample with a different seed, an ordinary model update of the kind that autonomous systems undergo routinely, only 39% of unconstrained and 37% of TRACE-CF recommendations remain valid. A customer who did exactly what they were told would, in most cases, still be declined by the updated model. The reason is

that minimal-cost counterfactuals sit, by construction, just across the decision boundary, where small changes in the model flip the outcome. Adding a validity margin of $\delta = 0.05$ raises robust validity to 65%, at a cost 43% higher than TRACE-CF without the margin and with 2.35 features changed.

This answers RQ3 and illustrates a trade-off that institutions must make explicitly: cheaper recourse is less reliable. TRACE exposes this trade-off through the margin parameter and requires that it be governed. A further implication is that a recourse promise should be time-bounded and tied to a model version, which the audit record makes possible.

6.5 E5: The conformal autonomy gate

Table 6.6 and Figure 6.5 report the behaviour of the gate.

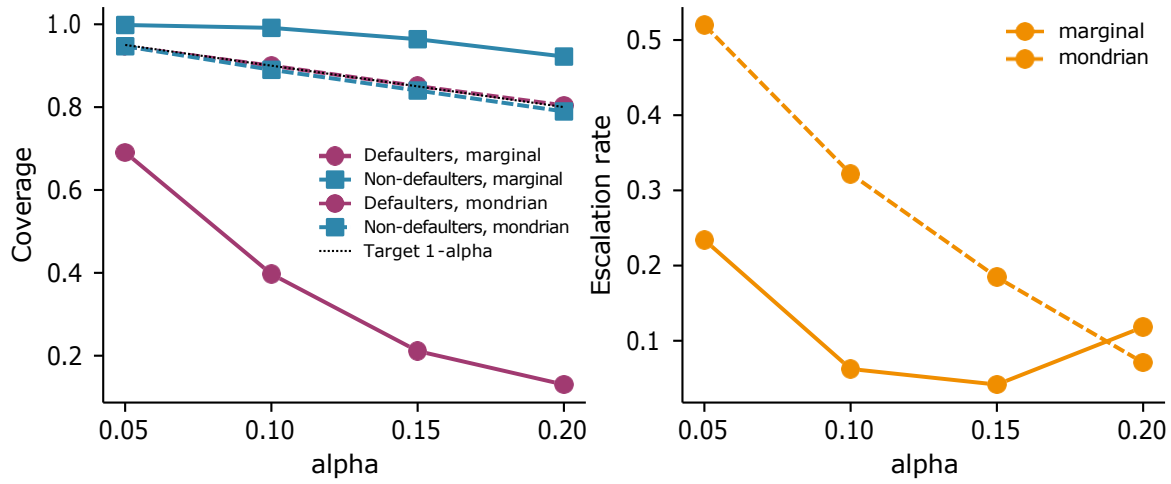

Figure 6.5: E5: Class-conditional coverage (left) and escalation rate (right) for marginal and Mondrian conformal prediction.

Marginal calibration hides a failure on defaulters. With marginal calibration at $\alpha = 0.10$, overall coverage is 0.907, satisfying the guarantee, and only 6.3% of cases are escalated. But coverage for actual defaulters is only 0.397. Because 86% of applicants do not default, the conformal threshold is set almost entirely by non-defaulters, and the procedure achieves its overall target by covering nearly all of them (0.991) while failing most defaulters. In autonomy terms, the gate confidently approves many applicants who go on to default, which is precisely the category of decision in which an institution would most want human oversight. At $\alpha = 0.20$, coverage of defaulters falls to 0.130.

Mondrian calibration restores the guarantee for each class. Class-conditional calibration achieves coverage close to the target for both classes (0.900 for defaulters and

0.890 for non-defaulters at $\alpha = 0.10$). The price is a higher escalation rate: 32.2% of cases at $\alpha = 0.10$ and 52.0% at $\alpha = 0.05$. The trade-off is direct: guarantees that matter for minority outcomes require more human review. Accuracy of autonomous decisions is lower under Mondrian calibration (0.840 against 0.901 at $\alpha = 0.10$), because the marginal procedure's high accuracy is largely the accuracy of approving non-defaulters.

Group coverage. Neither procedure guarantees coverage within protected groups. At $\alpha = 0.10$ with Mondrian calibration, coverage is 0.906 for group $A = 0$ and 0.864 for group $A = 1$. If coverage parity across groups is required, the gate should be calibrated within group–class cells, which is possible where the protected attribute is available for this purpose and lawful to use.

These results answer RQ4. The conformal gate gives a principled, distribution-free way of deciding when a system may act alone, but the choice of what the guarantee is conditional on is consequential and must be made deliberately. TRACE adopts Mondrian calibration by default for credit and requires that group-level coverage be reported.

6.6 E6: Detecting proxy discrimination with explanations

Table 6.7 and Figure 6.6 report the fairness audit.

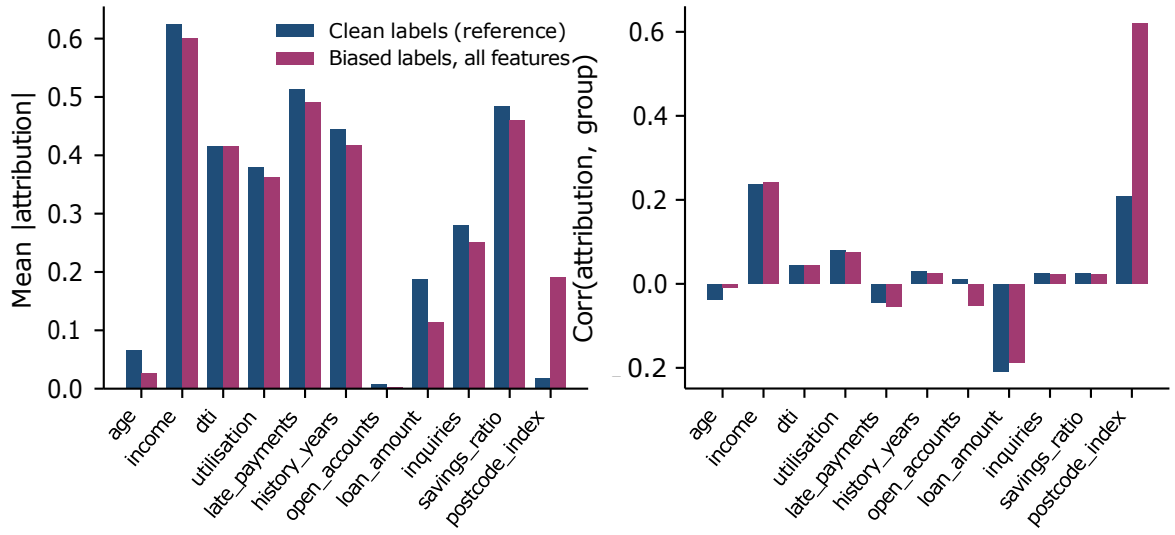

Figure 6.6: E6: Feature importance (left) and correlation of each feature's attribution with group membership (right), for models trained on clean and on biased labels.

With clean labels, the approval-rate gap between groups is 0.084. This gap is not discriminatory in the sense of the data-generating process: group $A = 1$ has lower income and higher debt on average, and these are genuine risk factors. The equal-opportunity

gap is essentially zero (0.0005), meaning truly low-risk applicants are approved at the same rate in both groups.

When the model is trained on biased labels, the approval gap doubles to 0.166 and the equal-opportunity gap rises to 0.024. The model cannot see the protected attribute, so the bias must travel through features correlated with it. The attribution audit identifies the route clearly. The importance of postcode_index, which has no causal role, rises more than elevenfold, from 0.017 to 0.191, and its attribution correlates with group membership at 0.62. No other feature combines material importance with such a high correlation; income, the next most group-correlated attribution (0.24), is a genuine risk factor whose correlation is the same under clean labels.

Removing the flagged proxy and retraining reduces the approval gap to 0.116 and the equal-opportunity gap to 0.008, with no loss of AUC against true outcomes (0.832 against 0.831). Part of the excess gap remains, because the bias in the labels also shifts how the model weights legitimate features that are correlated with group. Removing a proxy is therefore a partial remedy; correcting biased labels or reweighting training data would be needed to remove the remainder.

This experiment provides direct evidence, against a known bias mechanism, that attribution-based auditing can detect proxy discrimination, and it shows how the governance layer should respond: by flagging the feature, suppressing it from reason codes pending review, and assessing mitigation.

6.7 An end-to-end decision

To illustrate the framework's outputs for a single case, Figure 6.7 and Table 6.8 show the explanation and recourse for one declined applicant scored by the monotone HGB. The applicant's predicted default probability is 0.318, above the threshold of 0.25. The base log-odds of the background population is -2.20.

The four principal reasons are low income (24,550; +0.95), a high number of recent inquiries (4; +0.88), a low savings ratio (0.12; +0.66) and one late payment (+0.16). Two features work in the applicant's favour: a low debt-to-income ratio (-0.71) and a modest loan amount (-0.50). The reason codes generated by the governance layer are therefore:

1. Income insufficient relative to obligations.
2. Number of recent credit inquiries.
3. Level of savings.
4. Delinquency on a past account.

The recourse statement, computed with a validity margin of 0.05, is that the application

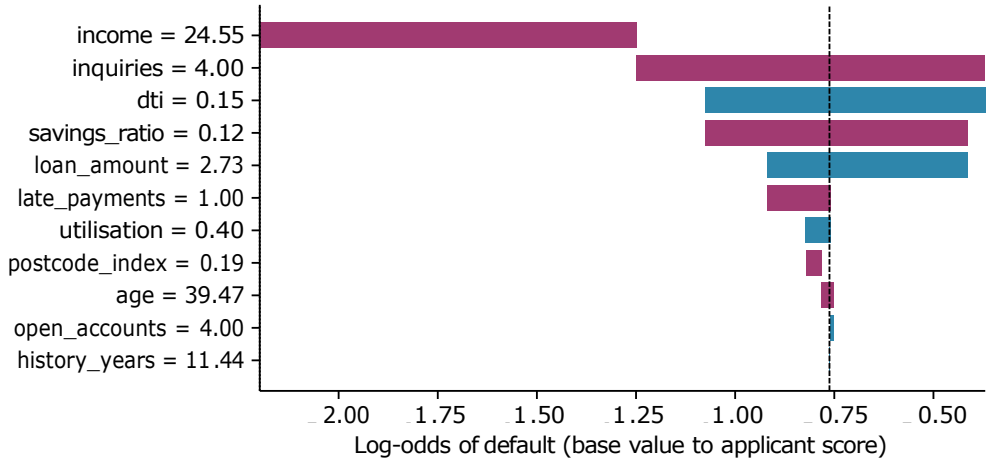

Figure 6.7: Explanation of one declined applicant: contributions of each feature to the log-odds of default, from the population base value to the applicant's score.

would be approved, with predicted default probability 0.195, if income were 29,080 (an increase of 18.5%) and credit utilisation were reduced from 40% to 31%. The search also reduced the loan amount by less than 0.1%, an immaterial change that a production system should remove by rounding recourse values to meaningful units.

Two observations follow. First, the recourse involves utilisation, which was not among the reasons for decline (its attribution was slightly favourable). Recourse and reasons answer different questions: reasons explain why the applicant is risky relative to the population, while recourse identifies the cheapest feasible path across the threshold. Both are needed, and conflating them, for example by advising customers to improve their top reason, can recommend changes that are infeasible (here, inquiries cannot be undone quickly) or unnecessarily costly. Second, the explanation is reproducible: given the model version, background version and inputs stored in the audit record, any reviewer or supervisor can regenerate it exactly.

6.8 Summary

The credit experiments show that constrained models lose nothing in accuracy (E1); that explanation error is dominated by the model rather than by good explainers, while LIME adds substantial error and instability (E2, E3); that recourse must be constrained to be actionable and margined to be robust (E4); that conformal gating is a principled autonomy mechanism but that marginal guarantees can conceal failure on the minority class (E5); and that attributions can reveal proxy discrimination against a known bias (E6).

Table 6.2: E2: Fidelity of explanations to the ground-truth attribution (200 applicants). Rank correlation with bootstrap 95% CI; relative ℓ_2 error is not reported for LIME, which explains probabilities rather than log-odds.

Model	Explainer	Rank corr. [95% CI]	Top-3	Sign	Rel. ℓ_2
Logistic regression	Exact Shapley (model)	0.960 [0.955, 0.964]	0.925	0.995	0.161
	KernelSHAP (200)	0.960 [0.955, 0.964]	0.927	0.995	0.161
	KernelSHAP (40)	0.953 [0.947, 0.959]	0.927	0.988	0.164
	LIME (1000)	0.862 [0.849, 0.873]	0.783	0.891	–
Gradient boosting	Exact Shapley (model)	0.870 [0.859, 0.880]	0.802	0.961	0.412
	TreeSHAP (interventional)	0.870 [0.859, 0.880]	0.802	0.961	0.412
	KernelSHAP (200)	0.870 [0.860, 0.879]	0.802	0.959	0.412
	KernelSHAP (40)	0.855 [0.840, 0.867]	0.803	0.945	0.418
	LIME (1000)	0.771 [0.757, 0.787]	0.748	0.884	–
Monotone HGB	Exact Shapley (model)	0.904 [0.895, 0.912]	0.815	0.983	0.291
	KernelSHAP (200)	0.904 [0.896, 0.912]	0.818	0.981	0.291
	KernelSHAP (40)	0.894 [0.884, 0.903]	0.828	0.965	0.297
	LIME (1000)	0.832 [0.819, 0.845]	0.767	0.891	–
MLP	Exact Shapley (model)	0.921 [0.914, 0.928]	0.882	1.000	0.203
	KernelSHAP (200)	0.920 [0.913, 0.927]	0.882	0.995	0.203
	KernelSHAP (40)	0.906 [0.897, 0.916]	0.885	0.973	0.212
	LIME (1000)	0.850 [0.837, 0.862]	0.778	0.891	–

Table 6.3: E2: Decomposition of explanation error. The model gap is $1-$ rank correlation between the model's exact attribution and the ground truth; the other columns give rank correlation between each explainer and the model's exact attribution.

Model	Model gap	TreeSHAP	KernelSHAP-200	KernelSHAP-40	LIME
Logistic regression	0.040	–	0.998	0.986	0.877
Gradient boosting	0.130	1.000	0.994	0.974	0.821
Monotone HGB	0.096	–	0.996	0.978	0.853
MLP	0.079	–	0.994	0.964	0.838

Table 6.4: E3: Stability of explanations (monotone HGB).

Explainer	Seed rank corr.	Seed top-3	Median rel. Lipschitz
LIME (1000 samples)	0.972	0.947	3.55
KernelSHAP (40 samples)	0.970	0.976	–
Exact Shapley	1.000	1.000	1.59

Table 6.5: E4: Recourse quality for 150 declined applicants.

Method	Valid	Cost	Sparsity	Non-action.	Plaus.	Robust
Unconstrained nearest CF	1.00	0.99	1.55	0.26	1.02	0.39
TRACE-CF (actionable)	1.00	1.12	1.95	0.00	1.01	0.37
TRACE-CF (actionable + margin 0.05)	1.00	1.60	2.35	0.00	1.01	0.65

Table 6.6: E5: Conformal autonomy gate on the test set. Coverage overall, for defaulters ($y = 1$), non-defaulters ($y = 0$) and by group; escalation rate; accuracy of autonomous decisions.

α	Calibration	Cov.	Cov. y=1	Cov. y=0	Cov. A=0	Cov. A=1	Escal.	Auto acc.
0.05	marginal	0.955	0.691	0.998	0.955	0.953	0.234	0.941
0.05	mondrian	0.947	0.946	0.947	0.953	0.935	0.520	0.889
0.1	marginal	0.907	0.397	0.991	0.914	0.893	0.062	0.901
0.1	mondrian	0.891	0.900	0.890	0.906	0.864	0.322	0.840
0.15	marginal	0.857	0.211	0.964	0.873	0.828	0.042	0.894
0.15	mondrian	0.842	0.851	0.840	0.861	0.806	0.184	0.806
0.2	marginal	0.810	0.130	0.922	0.836	0.761	0.118	0.918
0.2	mondrian	0.792	0.804	0.790	0.813	0.751	0.071	0.776

Table 6.7: E6: Group outcomes and proxy diagnostics. DP: demographic-parity gap in approval; EO: approval gap among truly low-risk applicants; Imp. and Corr.: mean absolute attribution of postcode_index and its correlation with group; AUC against outcomes drawn from the clean risk function.

Training condition	Appr. A=0	Appr. A=1	DP	EO	Imp.	Corr.	AUC
Clean labels (reference)	0.856	0.772	0.084	0.000	0.017	0.209	–
Biased labels, all features	0.825	0.659	0.166	0.024	0.191	0.620	0.831
Biased labels, postcode removed	0.806	0.689	0.116	0.008	0.000	0.000	0.832

Table 6.8: Audit-record excerpt for the declined applicant: values, recourse and exact Shapley attribution (log-odds).

Feature	Applicant value	Recourse value	Attribution
income	24.552	29.085	+0.950
inquiries	4.000	4.000	+0.879
dti	0.150	0.150	-0.707
savings_ratio	0.119	0.119	+0.662
loan_amount	2.727	2.726	-0.505
late_payments	1.000	1.000	+0.160
utilisation	0.397	0.308	-0.063
postcode_index	0.191	0.191	+0.039
age	39.471	39.471	+0.032
open_accounts	4.000	4.000	-0.009
history_years	11.440	11.440	-0.002

Chapter 7

Results II: Autonomous Portfolio Allocation

This chapter reports experiment E7, which instantiates TRACE for an autonomous agent that rebalances a portfolio every week. The emphasis is on RQ5: whether explanations of the agent's forecasts are adequate explanations of its decisions.

7.1 Investment performance

Table 7.1 and Figure 7.1 report out-of-sample performance over 787 weeks, net of transaction costs. The TRACE agent (labelled TRACE-PM) earns an annualised net return of 6.78% with volatility of 16.6%, a Sharpe ratio of 0.41, compared with 0.24 for equal weights and inverse-volatility weights and 0.27 for mean–variance with historical means. Its maximum drawdown is similar to that of the benchmarks, because all long-only portfolios are exposed to the same market factor.

Table 7.1: E7: Out-of-sample performance over 787 weeks, net of 10 bp transaction costs. Returns, volatility and drawdown in per cent; turnover is the mean weekly sum of absolute weight changes.

Strategy	Ann. return	Ann. vol.	Sharpe	Max DD	Turnover
TRACE-PM (GBM forecast + MV)	6.78	16.61	0.41	-52.1	0.619
Equal weight	3.84	16.00	0.24	-53.8	0.000
Inverse volatility	3.81	15.62	0.24	-53.1	0.006
Historical-mean MV	4.56	16.88	0.27	-46.2	0.102

The agent's advantage comes from the weak but genuine predictability built into the environment. It also comes at the price of very high turnover: on average 62% of the portfolio is traded each week, six times as much as historical-mean mean–variance. At 10 basis points per unit of turnover the agent remains ahead, but its advantage would be eroded at higher trading costs. This is not a property that a forecast-level explanation would reveal, but it is directly visible to a governance layer that monitors decisions, and in practice it would prompt the addition of a turnover penalty to the optimiser. The point for this thesis is not the performance of the agent, which depends entirely on the

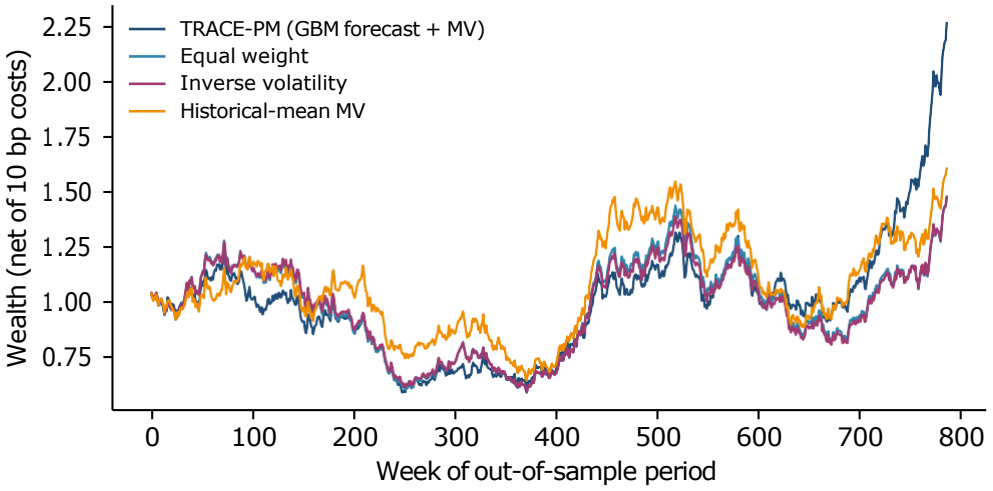

Figure 7.1: E7: Cumulative wealth of the strategies, net of transaction costs.

synthetic environment, but the relationship between its explanations at different levels.

7.2 Forecast-level explanations

Figure 7.2 (left) shows the share of forecast attribution carried by each feature at each annual retraining. Averaged over the sixteen retrainings, the credit spread carries the largest share (30.3%), followed by the valuation signal (24.4%) and 26-week momentum (16.0%); 4-week momentum, volatility and beta each carry about 10%. The shares drift from year to year, with the credit spread's share ranging from 17% to 47%, which is the kind of attribution drift that the monitoring component of the governance layer is designed to track.

The credit spread's prominence is not an error. In the environment, the credit spread genuinely predicts the market factor and therefore every asset's return; the forecaster correctly learns that expected returns are lower when spreads are wide. A risk manager reading forecast-level explanations would conclude that the agent is primarily a macro-timing system.

7.3 Decision-level explanations

The conclusion drawn from forecast-level explanations is wrong about the decisions. Figure 7.2 (right) shows the share of decision-level attribution, computed with the method of Chapter 4, averaged over all 787 decisions. The macro signal accounts for only 8.3% of the agent's allocation decisions. Valuation (35.3%), momentum (31.8%) and risk (24.6%) account for the rest.

The reason is structural. The credit spread shifts every asset's expected return by a

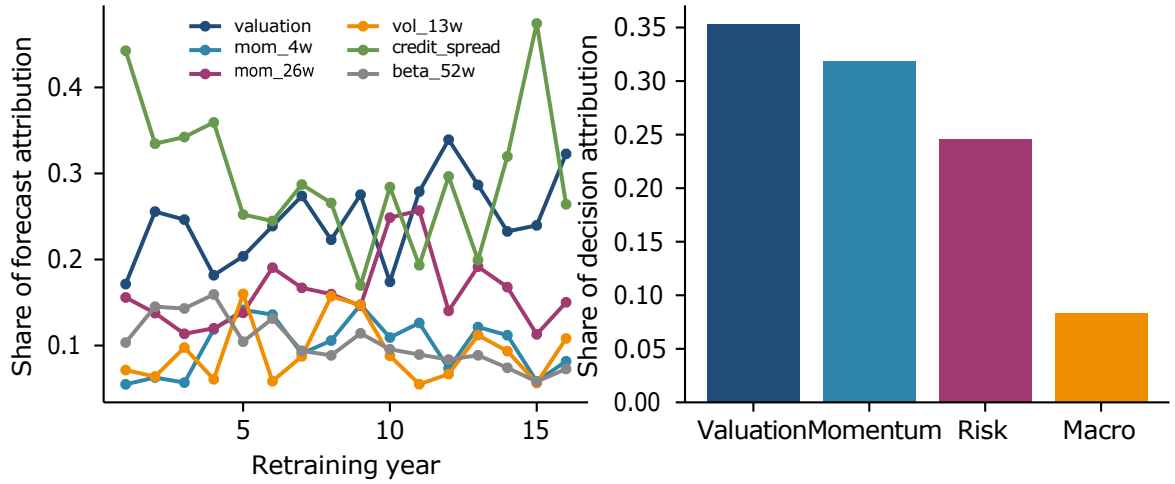

Figure 7.2: E7: Left, share of forecast attribution by feature at each annual retraining. Right, share of decision-level attribution by signal family, averaged over all weekly decisions.

similar amount, so it changes the level of $\hat{\mu}$ but hardly changes the *relative* attractiveness of assets. Because the portfolio is fully invested and long-only, a common shift in expected returns barely changes the optimal weights; it would matter only through the interaction with the covariance term and the constraints. What determines the weights are the signals that *differ across assets*: valuation, momentum and risk. The forecast-level explanation is accurate about the forecasts but misleading about the decisions.

The decision-level attributions satisfy efficiency to machine precision: across all decisions and assets, the largest discrepancy between the sum of group attributions and the change in weight relative to the reference portfolio was 1.1×10^{-16}.

7.4 Explaining a single rebalancing decision

Figure 7.3 explains the decision taken in week 200 of the out-of-sample period. The reference portfolio, the allocation the agent would choose if all signals were at their average values, is tilted towards low-beta assets: 23% in asset A1, 19% in A2 and 8% in A8. The actual decision puts the maximum 30% in each of A1, A3 and A8, eliminates A2 and A4 entirely, and holds 9% in A6 and 1% in A5.

The attribution gives a compact narrative that a portfolio manager or supervisor can check:

A8 rises from 8% to the 30% cap almost entirely because of momentum (+21.4 percentage points), with a small contribution from valuation (+4.4 points).

A3 rises from 17% to 30%, again driven by momentum (+19.8 points), partly offset by an unattractive valuation (-6.4 points).

A2 is eliminated, falling from 19% to zero, because of weak momentum (-24.1

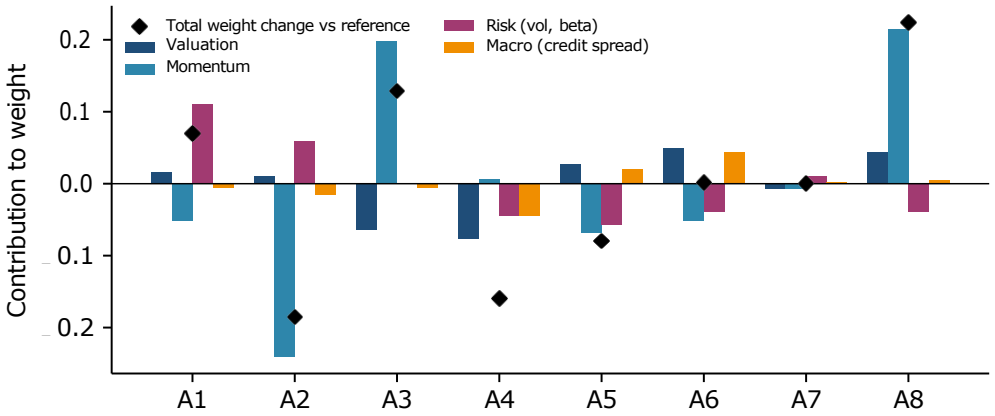

Figure 7.3: E7: Decision-level attribution for one weekly rebalancing. Bars show the contribution of each signal family to the change in each asset's weight relative to the reference portfolio; diamonds show the total change.

points), only partly offset by favourable risk characteristics (+5.9 points).

A1 rises from 23% to 30% because of its risk profile (+11.1 points), despite weak momentum (−5.1 points).

A4 is eliminated through a combination of valuation (−7.7 points), risk (−4.4 points) and the macro signal (−4.4 points).

In the language of the governance layer, this decision is a momentum-driven rotation into A3 and A8 out of A2, with a defensive overweight in A1. Such a statement can be logged, compared with the investment mandate (for instance, whether momentum concentration of this degree is permitted) and challenged.

7.5 Summary

E7 answers RQ5. In an optimiser-based autonomous agent, the drivers of predictions and the drivers of decisions can differ substantially. Here a signal carrying the largest share of forecast attribution (30%) carried only 8% of decision attribution. Explanations intended for accountability must therefore be computed at the level of the action. The decision-level Shapley attribution in TRACE achieves this exactly, is cheap when signals are grouped into a small number of families, and yields explanations that are meaningful to investment professionals.

Chapter 8

Discussion

This chapter interprets the findings in relation to the research questions and the literature, draws out implications for practice and regulation, and examines the threats to the validity of the conclusions.

8.1 Answers to the research questions

RQ1: Architecture. An explainable autonomous financial system needs more than an explainer bolted onto a model. The experiments show that the properties sought are interdependent. The choice of model (decision engine) changed explanation fidelity more than the choice of explainer did (E2). The usefulness of recourse (explanation engine) depended on constraints that encode domain knowledge and on a robustness margin that must be governed (E4). Whether a decision should be taken autonomously depended on uncertainty quantification whose conditioning had to be chosen with the class structure of the problem in mind (E5). And explanations served not only customers but the fairness audit (E6) and the oversight of an investment agent (E7). TRACE's four-layer structure reflects these dependencies by making assurance and governance first-class components with defined inputs and outputs.

RQ2: Fidelity. Exact or near-exact Shapley explainers are faithful to the model; LIME is substantially less so. The dominant source of error relative to the true drivers of risk is the model. This finding reframes the question of explanation quality. A long debate has focused on which explainer is best. The results suggest that, at least for Shapley-family methods with adequate budgets, practitioners should worry less about the explainer and more about whether the model has learned the right reasons, a question closer to Rudin (2019)'s argument than to the explainer-comparison literature. It also gives operational content to the observation that many models with similar accuracy can tell different stories: in E2, four models within 0.005 AUC of each other differed in top-3 reason accuracy by more than twelve percentage points.

RQ3: Recourse. Actionability constraints are essential, because unconstrained counterfactuals recommended impossible or discriminatory changes in a quarter of cases. Robustness is the harder problem: most minimal recourse did not survive a routine

retraining. The margin mechanism improved robustness substantially but at a price, confirming in a credit setting the concerns raised by Pawelczyk et al. (2020) and Upadhyay et al. (2021). The ethical implication is that a recourse statement is a promise, and an institution that makes it should either make it robust or qualify it.

RQ4: Autonomy. Conformal prediction provides a clean, assumption-light basis for deciding when a system may act. Its central weakness in credit, revealed in E5, is that the marginal guarantee is dominated by the majority class. A gate that satisfies its nominal 90% guarantee while covering only 40% of actual defaulters is formally correct and substantively misleading. Class-conditional calibration fixes this but more than quintuples the escalation rate at $\alpha = 0.10$. This makes the cost of meaningful human oversight explicit: it can be quantified and budgeted, and the level of autonomy becomes a governed parameter rather than an accident of model confidence.

RQ5: Decisions versus predictions. In the portfolio agent, a signal that dominated forecast explanations had little influence on decisions. This is not an artefact of the particular environment but a general consequence of decision rules that respond to relative rather than absolute predictions, of constraints that bind, and of optimisers that combine many predictions. Credit decisions that involve pricing, limit setting or portfolio-level constraints share this structure. Explanations for accountability should therefore be computed on the decision function.

8.2 Implications for practice

The findings suggest the following guidance for institutions building or validating autonomous financial decision systems.

1. **Validate reasons, not only explainers.** Checking that an explainer agrees with the model is necessary but insufficient. Where possible, test the full pipeline in a simulation with a known ground truth calibrated to the institution's own portfolio, so that the model gap can be estimated. Where it cannot be estimated, prefer model classes and constraints that reduce it.
2. **Prefer constrained models.** Monotonic constraints cost nothing in accuracy here, reduced the model gap by about a quarter and guarantee coherent reasons. They should be the default for credit.
3. **Do not use LIME for reason codes** without fidelity and stability evidence; in these experiments it changed principal reasons across reruns and was more sensitive to immaterial input changes.

4. **Version the background distribution** used for Shapley values, and treat it as part of the model.
5. **Separate reasons from recourse** in customer communications, and generate recourse only over actionable features.
6. **Govern the recourse margin** and tie recourse statements to a model version and validity period.
7. **Calibrate autonomy gates conditionally** on outcome class, and report coverage by protected group.
8. **Use attributions in fairness audits** to identify proxies, then evaluate mitigation by its effect on group outcomes rather than by removing features alone.
9. **Explain decisions where an optimiser is involved**, and monitor decision-level attribution drift as well as forecast-level drift.

8.3 Implications for regulation and supervision

The results bear on several regulatory questions. First, adverse-action rules require *specific and accurate* reasons. The model gap shows that reasons can be faithfully derived from a model and still be inaccurate as statements about why the applicant is risky. Supervisors may wish to ask institutions not only how reasons are generated but how the institution knows that the model's reasons are sound, and constrained models and simulation-based validation are concrete answers. Second, the human-oversight provisions of the EU AI Act require that oversight be effective. E5 shows that an apparently rigorous escalation mechanism can direct human attention away from exactly the cases that most need it. Supervisors assessing oversight arrangements should examine conditional, not only aggregate, performance of escalation rules. Third, recourse is not mandated in most regimes, but where institutions provide it the fragility demonstrated in E4 raises consumer-protection concerns. Finally, investment-management supervision has paid less attention to explanation than credit regulation has; E7 suggests that decision-level explanation is feasible and informative for the oversight of autonomous allocation.

8.4 Relation to the literature

The findings extend existing work in several directions. Studies of explainer disagreement and instability (Alvarez-Melis and Jaakkola, 2018; Slack et al., 2020) have shown that explanations can be unreliable relative to the model; this thesis adds a ground-truth

reference and shows that, for Shapley-family methods, reliability relative to the model is high while reliability relative to the world is limited by the model. Work on recourse robustness (Pawelczyk et al., 2020; Upadhyay et al., 2021) is confirmed in a credit setting with actionability constraints, and the margin mechanism is shown to offer a tunable compromise. The literature on conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2023) has noted the difference between marginal and conditional guarantees; E5 shows its practical magnitude when conformal sets are used to gate autonomy under credit-like class imbalance. Finally, applications of SHAP in finance (Bussmann et al., 2021; Bracke et al., 2019) have focused on predictions; the decision-level attribution introduced here addresses the decisions that autonomous systems actually take.

8.5 Threats to validity

8.5.1 External validity

All results are obtained on synthetic data. Real credit data contain missing values, measurement error, temporal drift, selection bias (outcomes are observed only for approved applicants) and dependencies far richer than those simulated. Real asset returns are less predictable, fat-tailed and non-stationary. The specific numbers reported, such as a model gap of 0.130 or a robust validity of 65%, should not be expected to hold elsewhere. The claims of the thesis concern mechanisms, and the mechanisms identified are general: model error that explainers faithfully reproduce, minimal recourse located near decision boundaries, majority-class dominance of marginal conformal guarantees, proxy transmission of label bias, and the difference between prediction and decision sensitivities. Each can and should be tested on real data, for which TRACE provides the instrumentation.

8.5.2 Internal validity

The ground-truth attribution depends on the interventional definition of the coalition value and on the background sample. Under an observational definition, attributions would be shared differently among correlated features (such as age and credit history), and conclusions about the model gap for those features could change. The model gap was measured with the same definition for the truth and for the models, so the comparison is internally consistent, and background sensitivity was quantified. Hyperparameters were set to reasonable defaults rather than tuned extensively; different settings could alter the ranking of models on fidelity, although not the qualitative finding that model gaps exceed explainer gaps for Shapley-family explainers.

8.5.3 Construct validity

The fidelity metrics capture whether explanations rank and sign features correctly; they do not capture whether explanations are useful to people. A rank correlation of 0.87 versus 0.96 may or may not change how a customer or reviewer acts. The recourse cost metric (MAD-normalised ℓ_1) is a proxy for the real effort of change, which differs greatly between, say, reducing utilisation and increasing income. The equal-opportunity gap uses a threshold on true risk to define "truly low-risk", which is a modelling choice.

8.5.4 Statistical conclusion validity

Fidelity results are averaged over 200 applicants with bootstrap confidence intervals, and the intervals for the different models do not overlap. Recourse results are based on 150 declined applicants and conformal results on 10,000 test cases. The portfolio experiment is a single simulated history; the Sharpe-ratio comparison should be read as illustrative, and the decision-level findings, which are structural, do not depend on it.

8.6 Ethical reflections

Explainability is sometimes presented as a solution to the ethical problems of automated decisions. The results counsel humility. An explanation can be faithful to a model that has learned the wrong reasons; recourse can be offered in good faith and fail; escalation can create the appearance of oversight without its substance; and removing a proxy can leave much of a bias in place. Explanations are best understood as instruments that make problems visible and contestable, not as guarantees that problems are absent. A framework such as TRACE is valuable to the extent that it obliges institutions to measure and govern these failure modes rather than to assume them away. Two further concerns deserve mention. Explanations can be gamed by applicants once reason codes are known, and by institutions that wish to present a model in a favourable light (Slack et al., 2020); fixed backgrounds, reproducible audit records and independent validation are partial defences. And the extensive use of a protected attribute for auditing, which E6 and E5 rely on, may itself be restricted in some jurisdictions; the tension between "fairness through unawareness" and the need to measure disparities remains unresolved in law and practice.

Chapter 9

Conclusion and Future Work

9.1 Summary of the thesis

This thesis set out to develop an explainable AI framework that allows financial decisions to be taken autonomously while remaining transparent, contestable, reliable and auditable. It began from three gaps in the state of the art: explanations are rarely verified against the true drivers of risk; explanation is rarely integrated with the other mechanisms an accountable autonomous system needs; and explanations target predictions rather than the decisions that autonomous systems take.

In response, the thesis developed TRACE, a four-layer framework comprising a decision engine with domain-constrained models, an explanation engine providing local Shapley attributions, reason codes, actionable and robust recourse and decision-level attributions, an assurance layer providing fidelity and stability monitoring, a conformal autonomy gate and fairness and proxy audits, and a governance layer producing customer communications, audit records, escalation workflows and model-risk reporting. It formalised decision-level attribution and the decomposition of explanation error into model and explainer gaps, and it evaluated the framework in synthetic credit and market environments in which the ground truth is known.

9.2 Principal findings

1. Domain-constrained models were as accurate as unconstrained ones: the monotone gradient-boosting model reached an AUC of 0.845 against 0.850 for the true model and 0.843 for an unconstrained ensemble.
2. Explanation error relative to the true drivers of risk was dominated by the model. Shapley-family explainers reproduced the model's exact attribution with rank correlations of at least 0.994, while the models' own attributions agreed with the truth at 0.870–0.960. Constraints reduced the model gap by about a quarter.
3. LIME added an explainer gap larger than any model gap and was more than twice as sensitive to immaterial input changes as exact Shapley attribution.
4. Unconstrained counterfactuals recommended non-actionable changes in 26% of

cases. Most minimal recourse failed after routine retraining (37–39% remained valid); a validity margin raised this to 65% at a 43% higher cost.

5. A marginally calibrated conformal autonomy gate met its overall 90% target while covering only 40% of defaulters. Class-conditional calibration restored coverage for each class at the cost of escalating 32% rather than 6% of cases.
6. Attribution-based auditing detected the proxy through which historical label bias entered the model, with the proxy's importance rising elevenfold and its attributions correlating 0.62 with group membership. Removing it reduced the approval gap from 0.166 to 0.116 without loss of accuracy, but did not eliminate the excess gap.
7. In an autonomous portfolio agent, the signal that carried 30% of forecast attribution carried only 8% of decision attribution, showing that explanations must be computed at the level of the decision.

9.3 Contributions

The thesis contributes (i) a reference architecture for explainable autonomous financial decision-making with defined components, interfaces and an audit-record schema; (ii) formal foundations comprising decision-level Shapley attribution with an efficiency guarantee, the model-gap/explainer-gap decomposition, and a conformal autonomy gate; (iii) a ground-truth evaluation methodology for financial XAI, including a synthetic credit environment with non-linear, interacting risk, a proxy and a controllable bias mechanism; (iv) empirical evidence on fidelity, stability, recourse, autonomy gating, proxy detection and decision-level explanation; and (v) practical and supervisory guidance derived from that evidence.

9.4 Limitations

The principal limitations are the reliance on synthetic data, which limits the transferability of specific numerical results; the absence of studies with human users, so that the usefulness of explanations to customers, reviewers and supervisors is argued rather than measured; the use of a single attribution definition (interventional Shapley); a recourse cost function that treats changes in different features as commensurable through statistical scale rather than real effort; and a single simulated market history for the portfolio experiment.

9.5 Future work

Several directions follow directly.

Validation on real portfolios. Applying TRACE to real credit and investment data, with simulation environments calibrated to those data to estimate model gaps, would test whether the mechanisms identified here operate at comparable magnitudes in practice.

Human-grounded evaluation. Controlled studies with loan applicants, credit reviewers and portfolio managers should examine whether TRACE's reason codes, recourse statements and decision-level narratives improve understanding, appropriate trust and the quality of human decisions on escalated cases.

Causal recourse. Recourse that respects causal relations among features, for example that reducing the loan amount changes affordability ratios, would make recommendations more realistic; combining causal models with the robustness margin is a natural extension.

Adaptive autonomy gates. Conformal procedures that adapt to distribution shift and that provide coverage within protected group–class cells, with formal control of the resulting escalation budget, would strengthen the assurance layer.

Decision-level explanation for sequential policies. Extending decision-level attribution from single-period optimisers to reinforcement-learning policies, where actions affect future states, is an open problem.

Adversarial robustness of the explanation pipeline. Testing TRACE against explanation-manipulation attacks of the kind demonstrated by Slack et al. (2020), and designing monitors that detect them, would harden the framework for supervisory use.

Explanations from generative AI. Language models are increasingly used to communicate decisions. TRACE's audit records provide a structured, verifiable basis from which natural-language explanations could be generated and checked for faithfulness.

9.6 Closing remarks

Autonomy and accountability are often presented as being in tension. The argument of this thesis is that they can be reconciled, but not by explanation alone. They can be reconciled by an architecture in which models are constrained by domain knowledge, explanations are verified rather than assumed, recourse is actionable and its fragility

governed, autonomy is granted case by case on the basis of calibrated uncertainty, fairness is audited through the same explanations offered to customers, and every decision leaves a record from which it can be reconstructed and challenged. The experiments reported here show both what such an architecture can achieve and where its limits lie. Those limits are themselves a finding: they mark the points at which human judgement remains indispensable.

Bibliography

Aas, K., Jullum, M. and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 103502.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B*, 82(4), 1059–1086.

Barredo Arrieta, A., D´ıaz-Rodr´ıguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56.

Bhatt, U., Xiang, A., Sharma, S., et al. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.

Bracke, P., Datta, A., Jung, C. and Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. *Bank of England Staff Working Paper* No. 816.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency (2011). *Supervisory Guidance on Model Risk Management* (SR Letter 11-7). Washington, DC.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, H., Janizek, J. D., Lundberg, S. and Lee, S.-I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.

Consumer Financial Protection Bureau (2022). *Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms*. Washington, DC.

Covert, I., Lundberg, S. and Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209), 1–90.

Dastile, X., Celik, T. and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263.

DeMiguel, V., Garlappi, L. and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5), 1915–1953.

Dixon, M. F., Halperin, I. and Bilokon, P. (2020). *Machine Learning in Finance: From Theory to Practice*. Cham: Springer.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 12 July 2024.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.

Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57.

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 3315–3323.

Heaton, J. B., Polson, N. G. and Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1), 3–12.

Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.

Hooker, S., Erhan, D., Kindermans, P.-J. and Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems 32*.

Janzing, D., Minorics, L. and Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2907–2916.

Jiang, Z., Xu, D. and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

Karimi, A.-H., Barthe, G., Schölkopf, B. and Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5), 1–29.

Khandani, A. E., Kim, A. J. and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 43:1–43:23.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C. and Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, 5491–5500.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4), 110–119.

Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Self-published.

Moody, J. and Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4), 875–889.

Mothilal, R. K., Sharma, A. and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

Parasuraman, R., Sheridan, T. B. and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A*, 30(3), 286–297.

Pawelczyk, M., Broelemann, K. and Kasneci, G. (2020). On counterfactual explanations under predictive multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 809–818.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77.

Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.

Ribeiro, M. T., Singh, S. and Guestrin, C. (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*, 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Sadinle, M., Lei, J. and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525), 223–234.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games II*, 307–317. Princeton: Princeton University Press.

Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647–665.

Sundararajan, M., Taly, A. and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

Upadhyay, S., Joshi, S. and Lakkaraju, H. (2021). Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems 34*, 16926–16937.

Ustun, B., Spangher, A. and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.

Verma, S., Dickerson, J. and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer.

Wachter, S., Mittelstadt, B. and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.

Appendix A

Data-Generating Process

A.1 Credit environment

For applicant i, with $U(a, b)$ the uniform, $N(\mu, \sigma)$ the normal, Beta(a, b) the beta and Pois(λ) the Poisson distribution:

$$
\begin{aligned}
A &\sim \text{Bernoulli}(0.35) \\
\text{age} &\sim U(21, 75) \\
\text{history_years} &= \min\{\max\{N(0.45(\text{age} - 21), 3), 0\}, \ \text{age} - 18\} \\
\text{income} &= \exp\{N(\ln 45 - 0.25A + 0.006(\text{age} - 40), \ 0.45)\} \quad \text{(thousands)} \\
\text{dti} &= \text{clip}(0.9\,\text{Beta}(2, 5) + 0.05A, \ 0.01, \ 0.95) \\
\text{utilisation} &\sim \text{Beta}(2, 3) \\
\text{late_payments} &\sim \text{Pois}(0.4 + 0.6\,\text{utilisation}) \\
\text{open_accounts} &\sim \text{Pois}(5) \\
\text{loan_amount} &= \text{income} \cdot U(0.1, 0.6) \\
\text{inquiries} &\sim \text{Pois}(1) \\
\text{savings_ratio} &\sim \text{Beta}(2, 6) \\
\text{postcode_index} &= 0.8A + N(0, 0.5) \\
Y &\sim \text{Bernoulli}\ \sigma(f^{\star}(x) + \beta A)
\end{aligned}
$$

with $f^{\star}$ given by equation (5.1), $\beta = 0$ in the clean environment and $\beta = 0.8$ in the biased environment. The clean environment uses seed 1; the biased and corresponding clean reference environments use seed 21.

A.2 Market environment

For weeks $t = 1, \ldots, 1100$ and assets $i = 1, \ldots, 8$, with seed 42:

$$
\begin{aligned}
cs_t &= 0.97\, cs_{t-1} + N(0,\, 0.15) \\
v_{i,t} &= 0.95\, v_{i,t-1} + N(0,\, 0.3) \\
F_t &= N(0.0012,\, 0.02) - 0.003\, cs_{t-1} \\
m_{i,t} &= \sum_{s=t-26}^{t-1} R_{i,s} \\
R_{i,t} &= \beta_i F_t + 0.0025\, v_{i,t-1} + 0.0015 \tanh(10\, m_{i,t-1}) + N(0,\, 0.025) \\
\tilde{v}_{i,t} &= v_{i,t} + N(0,\, 0.3)
\end{aligned}
$$

with β_i equally spaced between 0.6 and 1.4. The observed features at the end of week t are $\tilde{v}_{i,t}$, the trailing 4- and 26-week returns, the trailing 13-week standard deviation of returns, cs_t, and the 52-week rolling beta against the equally weighted average return.

Appendix B

Implementation Details

Table B.1: Hyperparameters and settings.

Component	Setting
Logistic regression	Standardised inputs; L-BFGS; maximum 2000 iterations; default ℓ_2 penalty
GBM	200 trees, depth 3, learning rate 0.05, seed 0
Monotone HGB	300 iterations, 15 leaves, learning rate 0.05, constraints $(0, -1, +1, +1, +1, -1, 0, +1, +1, -1, 0)$ in the feature order of Chapter 5, seed 0
MLP	Layers (64, 32), ReLU, $\ell_2 = 10^{-3}$, Adam, early stopping, maximum 300 epochs, seed 0
Shapley background	50 training applicants (seed 7 for E2, seed 8 for E3)
Fidelity sample	200 test applicants (seed 7)
KernelSHAP	200 and 40 coalition samples
LIME	1000 samples, no discretisation, reference statistics from 5000 training applicants, seeds 0–3
Decision threshold	$\tau = 0.25$
TRACE-CF	4000 candidates; 1–3 features per candidate; 20 candidates shrunk; three shrink passes with step 0.1; margin $\delta \in \{0, 0.05\}$
Retrained model (E4)	Monotone HGB on a bootstrap resample (seed 5), model seed 11
Conformal	Score $1 - \hat{\pi}_y(x)$; 6000 calibration applicants; $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$
Fairness audit	KernelSHAP (200 samples) on 300 test applicants; low-risk threshold: true default probability < 0.15
Portfolio forecaster	150 trees, depth 3, learning rate 0.05, subsample 0.8; expanding window; annual retraining from week 312
Portfolio optimiser	SLSQP; $\gamma = 5$; cap 0.30; long-only, fully invested; covariance from 104 weeks, 30% shrinkage to diagonal
Transaction cost	10 basis points per unit of turnover

Computation. On a single CPU core, exact Shapley enumeration (2048 coalitions, background of 50) for 200 applicants took 3 seconds for logistic regression, 11 seconds for the MLP, 17 seconds for the GBM and 74 seconds for the monotone HGB. For production use with more features, TreeSHAP or KernelSHAP with a budget validated by the fidelity monitor would replace enumeration.

Appendix C

Audit-Record Schema

Every decision taken or escalated by TRACE produces a record conforming to the schema below. The example is the declined applicant discussed in Chapter 6; numerical values are those produced in the experiment (income and loan amount in thousands).


```
{
  "record_id": "credit-000173",
  "timestamp_utc": "2026-11-26T09:14:03Z",
  "decision_type": "consumer_credit_application",
  "model": {"id": "monotone_hgb", "version": "1.0.0",
            "training_data_sha256": "
                d6f508ca08ab32725617cf4c1496edd456ca1b18668340ea2f7f5958973025ef",
            "monotone_constraints": [0,-1,1,1,1,-1,0,1,1,-1,0]},
  "explainer": {"method": "exact_interventional_shapley",
                "background_version": "bg-50-seed7", "output_space": "log_odds"},
  "inputs": {"age": 39.47, "income": 24.55, "dti": 0.150,
             "utilisation": 0.397, "late_payments": 1, "history_years": 11.44,
             "open_accounts": 4, "loan_amount": 2.727, "inquiries": 4,
             "savings_ratio": 0.119, "postcode_index": 0.191},
  "score": {"p_default": 0.318, "threshold": 0.25, "base_log_odds": -2.199},
  "autonomy_gate": {"method": "mondrian_split_conformal", "alpha": 0.10,
                    "q_hat": {"y0": 0.245, "y1": 0.926},
                    "prediction_set": [1], "route": "autonomous"},
  "attribution": {"income": 0.950, "inquiries": 0.879, "dti": -0.707,
                  "savings_ratio": 0.662, "loan_amount": -0.505,
                  "late_payments": 0.160, "utilisation": -0.063,
                  "postcode_index": 0.039, "age": 0.032,
                  "open_accounts": -0.009, "history_years": -0.002},
  "reason_codes": ["Income insufficient relative to obligations",
                   "Number of recent credit inquiries",
                   "Level of savings",
                   "Delinquency on a past account"],
  "recourse": {"method": "TRACE-CF", "margin": 0.05,
               "changes": {"income": 29.08, "utilisation": 0.308},
               "p_default_after": 0.195, "valid_for_model_version": "1.0.0"},
  "fairness_flags": [],
  "action": "decline", "actor": "system",
  "human_review": null
}
```

For this applicant the Mondrian gate at $\alpha = 0.10$ returned the singleton set $\{1\}$ (the non-conformity score for non-default, 0.318, exceeds its class threshold of 0.245, while that for default, 0.682, lies below its threshold of 0.926), so the decline was taken autonomously. The field training_data_sha256 holds the digest of the training data used to fit the model version; human_review records the reviewer identifier, decision and rationale for escalated cases. Records are append-only and are retained for the period required by the applicable record-keeping rules.