

Kevin Bauer
Andrej Gill
Katja Langenbucher
Lucia Franke

Explaining Explainability: On Credit Scoring, AI, and Consumer Agency

SAFE White Paper No. 116 | December 2025

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe

policy_center@safe-frankfurt.de | www.safe-frankfurt.de

Institutionalizing Explainability: On Credit Scoring, AI, and Consumer Agency*

*Kevin Bauer, Goethe University and SAFE
Andrej Gill, Gutenberg University Mainz and SAFE
Katja Langenbucher, Goethe University and SAFE
Lucia Franke, Goethe University*

December 2025

Abstract

The paper starts from a situation of information asymmetry on credit markets and zooms in on AI-enhanced credit scoring as an institutional response. It assumes the potential for expanding access to credit as well as the risk of discriminatory treatment of historically disadvantaged communities. Against this background, the paper explores legal requirements of „explainability“, using two recent European Court of Justice decisions as illustrations. The paper gives an overview of XAI methods along with their socio-technical and legal limits. It contributes to the discussion by suggesting to treat explanations as a public good and designing an intermediary institution which would act as a go-between connecting consumer data subjects and scoring companies.

I Introduction

In many markets, information asymmetries undermine allocative efficiency and market functioning (e.g., Stiglitz & Weiss, 1981). When one contracting party possesses substantially more information than the other, prices may be distorted, trust may erode, and welfare may decline. These concerns are especially salient in retail credit markets (Dobbie & Skiba, 2013). Consider an individual who applies for consumer credit in a foreign jurisdiction. The applicant has neither citizenship nor formal employment in that jurisdiction and lacks a local professional network and other conventional signals of creditworthiness. The bank faces a classic information asymmetry problem because it cannot readily assess whether the borrower is willing and able to repay the loan. This imbalance complicates risk assessment and pricing and can lead either to excessively cautious lending, which may produce above-market interest burdens and

* SAFE policy papers represent the authors' personal opinions and do not necessarily reflect the views of the Leibniz Institute for Financial Research SAFE or its staff.

credit rationing, or to overly optimistic lending, which may result in unsustainable leverage and borrower distress.

Credit scoring systems have emerged as a central institutional response to such informational imbalances. These systems transform observable characteristics and past behavior into a summary measure of creditworthiness. Traditional credit scoring relied on a relatively small set of variables, such as income, employment history, and repayment records (see, e.g., Gibbs et al. 2025). In the contemporary digital economy, however, lenders can draw on extensive behavioral, transactional, and digital trace data when forming predictions about credit risk (Berg et al., 2020).

These additional data sources create the possibility of expanding access to credit, particularly for underbanked groups that struggle to generate scores based on traditional inputs, for example, historically disadvantaged communities, recent immigrants, or refugees without an established credit history (Dobbie et al., 2020). Modern machine learning models can process high-dimensional information to produce credit scores that support more individualized lending decisions. In principle, finer-grained information permits a sharper distinction between relatively high-risk and low-risk borrowers and can therefore improve the allocative efficiency of credit markets, potentially yielding pricing that is more accurate for lenders, borrowers, and, by extension, financial supervisors.

At the same time, the use of complex machine learning models that exploit rich data environments introduces a distinct set of challenges, in particular with respect to the interpretability of scores. As the number of features describing individuals increases, flexible predictive models can better approximate the underlying data-generating process and capture patterns that generalize across contexts, yet this flexibility typically reduces transparency (Meske et al., 2022). Although such models often generate more accurate predictions, their internal logic is difficult to reconstruct in a manner that is understandable for human decision makers. As a consequence, it is hard to explain why a particular score was produced for a given applicant (Gramegna & Giudici, 2021). This limitation has important implications for multiple stakeholders in financial scoring, where decisions can have life-altering consequences for individuals, are subject to legal and ethical scrutiny, and, at scale, can affect the soundness of individual institutions and the stability of the financial system (Bastos & Matos, 2022).

When credit scoring systems are opaque, it becomes difficult to provide a meaningful justification for a specific score to the individual whose data is used, often referred to as the data subject. Concerns about fairness and accountability arise because the person affected cannot understand or contest the outcome. The user of the scoring model, typically the financial institution, is also constrained in its ability to determine whether a particular score is erroneous and to correct or override it. This situation sits uneasily

with legal rules in data protection and consumer credit law that require some form of explanation to consumers. It therefore creates challenges not only for lenders but also for financial supervisors (Bastos & Matos, 2022).

When a person is denied a loan, a simple reference to an algorithmic system that produced the outcome conflicts with established practices that govern interactions between credit scoring providers, banks, customers, and supervisory authorities. Affected individuals, regulators, and courts have traditionally insisted on an account of why credit was refused and how the institution's creditworthiness assessment operates (Citron & Pasquale, 2014; Sargeant, 2026). This traditional approach may be fundamentally altered if credit scoring increasingly relies on correlations embedded in hundreds of interacting variables that are difficult to interpret. Against this background, academic commentators and regulatory bodies worldwide have begun to examine how the fairness, lawfulness, and non-discriminatory character of creditworthiness assessments can be safeguarded (see, e.g., De Lange et al., 2022).

A recent case before the Court of Justice of the European Union illustrates these tensions between opacity, trade secrecy, and the right to an explanation. In February 2025, the Court delivered its judgment in Case C-203/22 concerning the automated credit assessment practices of Dun & Bradstreet Austria GmbH. An Austrian consumer had been refused the conclusion or extension of a mobile phone contract after the provider relied on a negative credit score supplied by Dun & Bradstreet. Although the consumer later obtained a score indicating very good credit standing, she could not receive a clear account of how the original score had been produced. When she sought access to more detailed information, Dun & Bradstreet relied on national trade secret rules to resist disclosure. The Court held that the combined provisions of Article 15(1)(h) and Article 22 of the GDPR require the controller to provide the data subject with all relevant information concerning the procedure and principles relating to the use of personal data with a view to obtaining, by automated means, a specific result, the obligation of transparency also requiring that that information be provided in a concise, transparent, intelligible and easily accessible form (Dun & Bradstreet Austria, 2025, para. 50). From a technical perspective, this indicates that the data subject should be given access to information about the model enabling them to understand and, where appropriate, contest the result; for example, this could in principle pertain even to the learned parameters of the model if this helps achieve the goal. Trade secrecy cannot be invoked in absolute terms to defeat this right. The judgment therefore signals that credit scoring systems that play a decisive role in contractual decisions must be accompanied by explanations that are understandable for non-specialists and tailored to the individual case.

Similar debates are unfolding in other jurisdictions. On September 23, 2025, the California Office of Administrative Law approved the California Privacy Protection Agency's (CPPA) regulations under the

California Consumer Privacy Act (CCPA) on automated decision-making technology. Beginning April 1st 2027, businesses using automated decision-making technology must conduct a risk assessment, provide pre-use notice to consumers along with an opt-out option, and provide the ability to appeal. Both the GDPR framework and the California rules therefore empower consumers to obtain information about the logic of automated credit and other significant decisions, even though they differ in scope, structure, and available remedies.

These horizontal data protection and privacy rules are complemented in the European Union by sector-specific legislation on consumer credit. Art. 18 para. 8 EU Consumer Credit Directive (EU) 2023/2225 (CCD) explicitly starts from the assumption that „artificial intelligence (AI) systems can be easily deployed in multiple sectors of the economy and society“. Following up on the GDPR, but asking for much more detail, the Directive explains that „the consumer should have the right to obtain a meaningful, comprehensive explanation of the assessment made and of the functioning of the automated processing used, including the main variables, the logic and risks involved, as well as the right to express the consumer’s point of view and to request a review of the assessment of the creditworthiness and a review of the decision on whether to grant credit“. Article 18 para. (8) CCD lays down the details: „where the creditworthiness assessment involves the use of automated processing of personal data, Member States shall ensure that the consumer has the right to: (a) request and obtain from the creditor human intervention, consisting of the right to request and obtain from the creditor a clear and comprehensible explanation of the assessment of creditworthiness, including on the logic and risks involved in the automated process of personal data as well as its significance and effects on the decision“.

Explaining to consumers how AI systems are employed is at the heart of the AI Act, too. The scope of the AI Act, mostly, concerns developers and deployers of AI systems. This explains why it requires AI systems to be “developed and used in a way that allows (...) explainability”, including several rules to require manual-like instructions. Despite the AI Act’s focus on developers and deployers, explainability with the aim to achieve transparency about how AI shapes decision-making processes was considered important enough to justify the sole rule in the entire AI Act that focuses on consumers. “Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system [that is a high-risk AI system]“, Art. 86 AI Act reads, „which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken“.

Note that the text of the AI Act zooms in on the role of the AI system in the decision-making procedure. This is to be understood against the AI Act repeatedly stressing that other Union rights, especially the

GDPR, remain unaffected. The AI Act does not explain how the “role of the AI system in the decision-making procedure” differs from the GDPR’s “meaningful information about the logic involved”. One plausible interpretation is that the GDPR establishes an overarching principle that focuses on data protection, profiling, and the use of sensitive data, whereas the AI Act adds a residual safeguard that aims to ensure transparency whenever AI systems influence decision-making in ways that might otherwise escape scrutiny.

These provisions raise two sets of questions. First, what options exist to render scoring models interpretable, and do these options comply with requirements in current regulations. Second, how should explanations of scoring methods be organized so that they secure lawful, incentive-compatible, and truthful disclosure that contributes to the protection of consumers.

The essay at hand aims to address these questions. We survey methods that can be used to render scoring models interpretable and clarify what these explainability measures can realistically deliver. Thereby, we adopt socio-technical, economic and legal perspectives to examine the consequences of these methods and assess whether they are compatible with the emerging regulatory framework. Based on the presented insights, we discuss the idea for an explainability intermediary, in which a trusted entity produces explainability reports, in order to mitigate the obstacles that currently prevent explanations from achieving their intended protective function.

II Why do we need explanations?

From the perspective of the consumer, the variables used in a scoring model and the logic applied to arrive at an AI-based prediction of creditworthiness are of particular interest. At the same time, making these aspects understandable constitutes a sophisticated socio-technical challenge. Explanations of why an AI model produces a specific prediction are not merely technical summaries produced in isolation. They are created, communicated, and interpreted by human actors who have specific motivations for providing and receiving explanations (Miller et al., 2019).

Different stakeholders have different reasons to care about understanding why an AI model generates a particular credit score in addition to achieving high predictive performance (Martens et al., 2025). During model development, developers need to understand why the model behaves in a certain way in order to detect and correct problematic behavior, for example, when the model assigns implausible weights to particular inputs. In a well-known example from image classification, a model relied on snow in the background rather than on image regions that showed the animals to distinguish between wolves and dogs, a behavior that would have remained unnoticed without interpretability methods (Ribeiro et al., 2016).

Banks, as providers and users of explanations, must be able to justify to consumers how a prediction was reached on the basis of the available input information, both for reasons of legal compliance and in order to support accountability and trust. Explanations can also enable human decision makers in the loop to assess whether the reasons for a specific credit score are plausible, which can inform decisions about when to follow or overrule the model. Scored consumers can benefit from explanations in several ways (Rosenfeld & Richardson, 2019). Explanations may help individuals to assess whether legal action is warranted, for example, when they suspect discriminatory practices. They can also reveal actionable insights into how a person might improve their credit score, such as by reducing certain recurring expenditures or adjusting other aspects of their financial profile. In this sense, explanations are not only a mechanism for contestation but also a tool for learning and adaptation on the side of the consumer (Bauer et al., 2021).

On a broader level, the availability of explanations can foster social acceptance of AI-based credit scoring. A widely held assumption is that humans find it easier to trust a system that offers reasons for its outcomes than a system that remains a black box. From this perspective, explainability is relevant not only for individual fairness but also for the perceived legitimacy of algorithmic credit decisions in society (Bauer et al., 2021). At the same time, this assumption presupposes that meaningful and useful information can in fact be provided. This raises a difficult question for regulators and firms. Will consumers be satisfied with explanations that are technically accurate yet generic and hard to understand, or should the use of opaque systems be restricted even when their predictive performance clearly exceeds that of more transparent models?

Explainability also has implications for competition in credit markets. Where several lenders rely on algorithmic scoring, the capacity to communicate intelligible reasons for credit decisions can serve as a quality dimension of financial products. Transparent models can strengthen consumer trust and reduce search and switching costs, which may intensify competition on fairness, accuracy, and service quality rather than only on price (Heidhues et al., 2017). By contrast, opacity can reinforce market power. Providers that operate proprietary black box systems can shield themselves from scrutiny, making it difficult for consumers, regulators, and competitors to evaluate performance or challenge discriminatory outcomes. Over time, this may create informational lock-in because those who control large, non-transparent models accumulate advantages that stem less from efficiency gains than from the inability of others to interrogate or replicate their scoring logic (Berg et al., 2020).

Existing regulatory approaches only partially address these competitive dynamics. The AI Act contains transparency and documentation duties, although these largely govern the relationship between developers, deployers, and supervisory authorities (Regulation (EU) 2024/1689). The Act pays

comparatively little attention to consumer protection and does not engage with competition law concerns. As a result, statutory obligations under the AI Act are unlikely to determine how explainability is used in competitive positioning. A different picture emerges under data protection and consumer credit law. These frameworks do not require public disclosure in a strict sense. However, credit scoring firms and banks that are confronted with individual requests to explain how they score will know that whatever they reveal to a customer is unlikely to remain confidential for long. This possibility can influence both the design of scoring systems and the willingness of institutions to rely on highly opaque models.

Against this background, it is crucial to understand how legal frameworks in data protection, consumer credit, and AI regulation conceptualize the explanations or information that must be given to a consumer. The same question arises in the more complex relationship between a bank's internal rating specialists and its banking supervisor, where explanations serve as a basis for oversight, challenge, and potential intervention (European Central Bank, 2025).

III What constitutes a good explanation

A common goal of explanations about why a model behaves in a particular way is to provide humans with an understanding of the cause of a prediction or decision (Doshi-Velez & Kim, 2017). From this perspective, a good explanation helps to answer a "why" question. Psychologically, answering such questions can satisfy basic needs to reduce uncertainty and to feel competent (Miller, 2019). Explanations become especially important when outcomes are unexpected and conflict with a person's prior knowledge, past experience, or projections, because they help individuals to reconcile perceived inconsistencies (v. Zahn et al., 2025).

In the context of credit scoring, an explanation would aim to give involved stakeholders an understanding of why the model produces certain scores based on the inputs. This can happen on two different levels: a local and a global level (Bauer et al., 2021). A local explanation clarifies the reasons for an individual consumer's score. If a low credit score is entirely unexpected, a local explanation can reduce the perceived gap between the anticipated and the actual outcome. It can provide a basis for legal recourse when the explanation reveals an unlawful reason for the discrepancy and can offer guidance on how to adjust individual behavior in order to improve future scores. From the perspective of the individual bank client, such an explanation is often the only way to assess whether the outcome reflects their own financial conduct or instead suggests an error or structural bias in the scoring system. A global explanation, by contrast, does not focus on a single individual but reveals the average behavior of the model across many cases. Global explanations are particularly valuable for developers, who use them to align expectations with actual model performance, and for banking supervisors, who can rely on them to detect unexpected

aggregate patterns that may indicate discriminatory treatment based on protected characteristics and may therefore call for supervisory action.

Local and global explanations respectively answer the question why a model produces a specific output for a given input and why the model behaves the way it does overall (Molnar, 2020). Interestingly, from its ruling it seems that it is this kind of local, counterfactual explanation that the European Court of Justice in *Dun & Bradstreet* had in mind. Put differently, the European Court of Justice suggested an approach that allows for a form of causal reasoning of why this specific individual received the very score she did instead of another one, enabling her to learn what she could change in order to receive a different prediction. While this ruling seems plausible given our previous considerations, it is pivotal to gauge the feasibility of such a form of explanation both from a purely technical (methods to generate explanations) and a social perspective (design of explanation provision mechanisms).

IV eXplainable AI

With the rapid adoption of complex machine learning based systems in decision making, research in computer science and information systems has increasingly focused on eXplainable AI (XAI). XAI refers to methods that help humans understand why AI models produce specific outputs (Bauer et al., 2021). In credit scoring, these methods are often presented as tools to increase transparency and accountability. However, what computer scientists aim to capture with explainability techniques frequently differs from what consumers, regulators, supervisors, and legislators expect to be explained (Martens et al., 2025). This gap has led to persistent misconceptions about which aspects of model behavior can be meaningfully communicated and to whom.

IV.1 White-box and Black-box models

There are different ways to make the logic behind credit scoring models accessible to human understanding. One approach is to rely on white box models that are inherently interpretable because of their simple structure and relatively low mathematical complexity. Linear and logistic regression are standard examples, where learned patterns are encoded in coefficients of additive terms (e.g., Wang et al., 2015). In such models, the influence of an input feature such as annual income on a credit score prediction can be read directly from its coefficient. For example, if the coefficient for age in a linear regression for credit scoring is equal to 2, then the credit score increases, *ceteris paribus*, by two units per unit of age. Similarly, a simple decision tree can be interpreted by following the sequence of learned decision rules that place a new applicant into a group of similar individuals whose creditworthiness is historically known (Szwabe & Misiorek, 2018). For example, if an applicant is predicted to have a credit

score of 700, this means that the average historical credit score of historical applicants that are similar according to certain features has been 700, whereby similarity can be read from the logical rule that the decision learned.

This inherent interpretability supports several objectives at once. It allows banks to justify predictions, facilitates the identification and correction of model errors, and enables developers and domain experts to check whether the model is consistent with substantive knowledge and expectations. These properties are especially valuable in regulated domains such as credit scoring, where transparency and accountability are central legal and supervisory concerns. The main limitation of inherently interpretable white box models is their restricted capacity to capture complex, non-linear relationships in high-dimensional data. Because their structure is constrained, they may fail to detect intricate patterns that would improve predictive accuracy. In practice, this often leads to a trade off between interpretability and performance when the underlying data-generating process involves many interacting variables and latent factors (Molnar, 2020).

In theory, simple models do not necessarily perform worse than complex models on new, unseen data. If the true latent mechanism that determines creditworthiness is a simple function of income and years of education, a transparent logistic regression could exhibit better out-of-sample performance than a neural network or a random forest. In real-world credit markets, however, the processes that models attempt to approximate are unlikely to be so simple. As a result, non-interpretable models frequently achieve higher predictive performance. Choosing an inherently interpretable model in such settings, therefore, often means accepting a loss of potential predictive accuracy in exchange for greater transparency (e.g., Rudin, 2019).

From an economic and financial perspective, the trade-off between interpretability and predictive performance has important implications for credit allocation, inclusion, and systemic stability (Fuster et al., 2022). Simpler, inherently interpretable models such as logistic regressions or decision trees offer clear benefits for auditability, governance, and consumer understanding. They allow lenders, regulators, and consumers to trace how inputs translate into outcomes and to identify potential sources of bias or error. Because these models capture only limited functional relationships, however, they often produce coarse classifications of credit risk. Borrowers with very different true risk profiles may receive similar scores, which leads to mispricing and inefficient capital allocation. For lenders, this can result in conservative lending policies and average pricing of risk heterogeneous portfolios. New entrants, self employed borrowers, and consumers without extensive credit histories are particularly affected, which may intensify credit supply contractions during downturns when institutions tighten standards in response to rising uncertainty about defaults.

By contrast, complex and data-intensive models can detect subtle non-linear patterns and interactions across many variables, capturing predictive signals that traditional models miss. This can improve risk-based pricing and expand access to credit, especially for previously underserved groups such as young borrowers or migrants with limited histories. These efficiency gains come with new costs. High-dimensional models are difficult to interpret and validate, which complicates assessments of fairness, stability, and causal soundness (e.g., Doshi-Velez & Kim, 2017). Algorithms that are formally race blind or demographically neutral can still produce persistent disparities through correlations with unobserved or proxy variables, so that higher accuracy does not guarantee fair treatment (Kordzadeh & Ghasemaghaei, 2022). The predictions of such models may also be sensitive to small changes in input data or to shifts in the macroeconomic environment. When these systems are widely adopted, correlated errors and common training data can contribute to model fragility, procyclical lending patterns, and unintended amplification of systemic risk.

For individual borrowers and small and medium-sized enterprises (SMEs), the choice between lenders that leverage interpretable white box models and complex black box models can determine whether they gain or lose access to credit. Individuals with short or thin credit files, such as young adults, recent migrants, or those who mainly use cash, are often poorly represented in traditional scoring systems that rely on long repayment histories or stable employment. They may benefit from complex models that draw on alternative data, yet the opacity of these models makes it hard to see which aspects of financial behavior can meaningfully improve creditworthiness (see Frost et al., 2019). Interpretable models are easier to understand and respond to but often fail to capture compensating information that would allow lenders to identify low-risk borrowers in these groups. A similar tension exists for SMEs, which frequently lack standardized, long-term financial records. Richer, more flexible models can better capture sector-specific patterns and qualitative strengths, but offer little transparency when credit is denied or priced unfavorably. As a result, both individuals and SMEs face a structural paradox. The models most capable of identifying their idiosyncratic strengths are often least able to explain themselves, whereas the models that offer clear explanations are often least capable of recognizing their creditworthiness.

To address the opacity of complex machine learning models, which are well-suited to learning intricate patterns, researchers in information systems and computer science have developed techniques that generate explanations after model training. These post hoc explanation methods decouple model construction from model interpretation and are intended to make it possible to use high-performing models while still providing some level of interpretability. In practice, they work by placing explanatory tools on top of an already trained model and then producing local or global insights, depending on the method. This modular approach appears to offer flexibility in model selection in domains where

interpretability is a regulatory or practical requirement, because the explanatory layer can be added after the predictive model is finalized (e.g., Ribeiro et al., 2016b). At the same time, regulators and legislators need to recognize that such explanations remain approximations produced by auxiliary tools.

The repertoire of post hoc methods continues to grow and offers different ways to interpret model behavior. Scholbeck et al. (2019) observe that many widely used approaches, especially those that are model agnostic, follow the SIPA framework: they **S**ample data points, **I**ntervene by manipulating input features, obtain **P**redictions for these altered inputs, and **A**ggregate the results to form an explanation. This general procedure underlies several prominent techniques. In what follows, we concentrate on two widely used approaches that are representative of a broader and influential class of methods. The first is feature attribution, for which we consider SHAP, and the second is counterfactual explanation, for which we consider DiCE.

IV.2 A feature attribution method: SHAP

Feature attribution methods estimate how individual input features contribute to a model's prediction by expressing the output as a sum or combination of feature-specific contributions. In this way, they decompose complex model predictions into contributions that can be associated with particular inputs. Such methods are especially useful for assessing whether certain inputs influence model outputs in unexpected or problematic ways. They provide users of model predictions with additional information for deciding when to overrule an automated output and offer developers insights that can support model improvement. Prominent approaches include local feature attribution methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), as well as related techniques for assessing feature influence such as partial dependence plots and permutation feature importance, with SHAP being particularly widely used in research and practice.

SHapley Additive exPlanations (SHAP) build on ideas from coalitional game theory to assign contributions to individual feature values. Each feature value is treated as a player in a game and the model prediction is the payout. SHAP evaluates the marginal contribution of each feature value across many possible subsets of features in order to capture non-linear interactions (Lundberg & Lee, 2017). The result is a set of feature-specific values that explain why a given prediction differs from the average prediction for the population. By construction, the sum of all SHAP values for an instance together with the average prediction equals the actual prediction. In this sense, SHAP provides a contrastive explanation that expresses a complex prediction as a linear surrogate that approximates the behavior of the underlying model.

SHAP explanations are typically local, since they pertain to single instances. Aggregating SHAP values across many cases can reveal global patterns in model behavior, for example, systematic over-reliance on particular variables, which is useful for model validation and regulatory scrutiny. It is important to note that SHAP values are not derivatives. A positive SHAP value does not mean that increasing the corresponding feature will increase the prediction (Lundberg & Lee, 2017). Instead, it reflects the contribution of the observed feature value within the full context of all other features.

Consider a simple example. Suppose a credit scoring model has an average predicted score of 650. For applicant Alice, the model predicts a score of 720. SHAP attributes the seventy-point difference to the following feature contributions. High annual income contributes plus thirty, a low debt to income ratio contributes plus twenty, an absence of late payments contributes plus twenty-five, a long credit history contributes plus ten, a recently opened new account contributes minus ten, and high credit utilization contributes minus five. The SHAP values sum to seventy, and 650 plus 70 equals 720, which reconstructs Alice's prediction. Each SHAP value is computed in light of Alice's complete feature profile. A change in one input, even a small change, would in principle alter the full set of SHAP values rather than only the value for that particular feature. As a consequence, the contribution associated with a feature such as gender may differ for Alice and for another applicant, such as Barbara, which makes direct inferences about discrimination at the individual level difficult. Only by aggregating SHAP values across many applicants is it possible to obtain more systematic insights into how features influence predictions on average.

IV.3 A counterfactual explanation method: DiCE

Counterfactual explanations form a core class of explanation methods. They show how a model's prediction would change if one or more input features were different. The typical goal is to identify small changes to the input that would lead to a different predicted outcome, thereby offering actionable insights to end users (e.g., Dandl et al., 2020). In practice, these methods operate by modifying the feature values of an instance, querying the model for the new prediction, and focusing on those modifications that produce a meaningful change, for example, a shift from credit rejected to accepted or a score crossing a specified threshold. A counterfactual explanation, therefore, describes a change to the input that is intended to be as small as possible while achieving a predefined target output (Molnar, 2020).

These explanations are often regarded as user-friendly because they are contrastive and selective and usually involve changes to only a few features. A technical advantage is that they can be implemented in a model-agnostic way. In such settings the method does not necessarily require access to the training data or to model internals but only to the model's prediction function, which can be accessed through an

interface such as a web API. Counterfactuals are particularly well-suited to clarifying why a specific prediction was produced instead of another possible outcome. Counterfactual approaches construct new data points for which the model's output can be observed directly. At the same time, counterfactuals are essentially local. They do not readily aggregate into a summary of global model behavior and their primary value lies in illustrating how one particular decision could have differed under slightly altered conditions (Molnar, 2020).

DiCE, Diverse Counterfactual Explanations (Mothilal et al., 2020), is a widely used implementation of this approach. It addresses a common limitation of standard counterfactual methods, namely their tendency to produce only a single or a very narrow set of alternatives. DiCE generates multiple counterfactuals that all achieve the target outcome but differ along various feature dimensions. It formulates counterfactual generation as an optimization problem that balances three main criteria. Proximity measures how close a counterfactual is to the original input, diversity measures how different the counterfactuals are from each other, and validity requires that each proposed counterfactual actually leads the model to output the desired prediction (Mothilal et al., 2020). Intuitively, DiCE maps out several plausible alternative scenarios in which the model would have made a different decision, giving users a set of feasible options rather than a single prescription.

Consider an example in the context of a credit scoring model. Suppose the model predicts a score of 640 for applicant Bob, which is below the threshold for loan approval. DiCE may generate several counterfactual profiles that all yield a score above 700. One counterfactual could suggest increasing annual income from 40,000 to 60,000 while holding other features constant. Another might keep income unchanged but reduce credit utilization from 85 percent to 30 percent. A third could recommend reducing the number of recently opened accounts and increasing the length of credit history. These counterfactuals illustrate different ways in which the outcome could have changed. From the perspective of detecting systematic problems in the model, however, counterfactual explanations for isolated cases are limited because they provide only local insight and do not reveal structural patterns of misbehavior on their own.

IV.4 Socio-technical limitations

Feature attribution and counterfactual explanation methods both follow the SIPA logic of sampling, intervening on inputs, obtaining predictions, and aggregating results (Scholbeck et al., 2019), yet they differ in their assumptions, computational procedures, and the types of insight they provide. As a result, they can yield different explanations for the same model and even for the same data point. From a socio-technical perspective, this plurality of explanations creates discretion in the choice of method and representation. It opens room for strategic selection of explanations, intentional or unintentional misuse,

and new forms of bias that interact with human cognitive limitations such as confirmation bias (e.g., Lakkaraju & Bastani, 2020). These dynamics help to explain why the use of explainability tools has so far produced mixed empirical results with respect to improvements in human AI collaboration (e.g., Poursabzi-Sangdeh et al., 2021; Bauer et al., 2023).

Adding complexity to the issue, this variability does not only arise from the choice of method but also from deeper methodological and contextual dependencies, which raise concerns about reliability and interpretability (e.g., Fernandez-Loria et al., 2022). Many post hoc methods, including SHAP and DiCE, are sensitive to user-defined inputs and modelling choices. In SHAP, for example, the background dataset used as a reference distribution has a strong influence on the resulting attributions. Different selections of this background data can lead to different decompositions of the same prediction. This flexibility creates scope for misleading explanations, whether deliberate or inadvertent. A model developer or institution could, for instance, choose background data that reduces the apparent influence of a sensitive feature and thereby masks discriminatory effects. In the Alice example discussed above, where the model average is 650, and her score is 720, different background datasets could produce different SHAP value decompositions of the seventy-point difference and might conceal the extent to which her gender reduces her score, both in her case and on average across applicants. Sampling problems and other technical issues can similarly generate distorted explanations unintentionally, without any actor being aware of the distortion.

For counterfactual explanations, the problem of multiplicity is even more pronounced. Many different counterfactuals can exist for a single case, sometimes referred to as the Rashomon effect (e.g., Müller et al., 2023). DiCE, which is designed to generate diverse counterfactuals that all attain the same target outcome, illustrates this phenomenon clearly. Each counterfactual corresponds to a different hypothetical set of changes in the inputs, yet all of them induce the same decision by the model (Molnar, 2020). When the interests of explanation providers and recipients diverge, this ambiguity becomes critical. In credit scoring, a bank might highlight counterfactuals that place responsibility on the applicant, such as higher income or a longer credit history, while de-emphasizing counterfactuals that reveal illegal patterns that actually drive the rejection.

These issues affect stakeholders in different ways. For users who rely on model predictions to inform decisions, ambiguity in explanations reduces their practical value for contesting or overruling adverse outcomes. When explanations can be selectively framed or adjusted, users cannot easily determine whether an unfavorable result reflects an accurate model judgment or a systematic error. In such settings, interpretability does not reliably translate into accountability or effective oversight.

For developers, instability across explanation methods complicates model debugging and improvement. When two widely accepted post hoc techniques attribute a prediction to different features, it becomes difficult to distinguish genuine properties of the model from artifacts of the explanatory method. This uncertainty can misdirect optimization efforts towards correcting perceived problems that originate in the explanation procedure rather than in the model itself.

For data subjects, inconsistencies and degrees of freedom in constructing explanations directly threaten fairness and transparency. Individuals who receive a justification for a specific prediction expect a clear and truthful account of why the decision was made (e.g., Wachter et al., 2018). If multiple, equally plausible explanations exist for the same outcome, or if explanations depend on choices about background data that are not visible to the data subject, this expectation is undermined (Lakkaraju & Bastani, 2020). People cannot readily distinguish between faithful and selectively framed explanations, which can erode trust in both the model and the institution using it.

Against this background, one can argue that post hoc explanation methods such as SHAP and DICE introduce an additional layer of complexity to the deployment of powerful machine learning models. They can provide interpretable summaries to some extent, provided that recipients understand their limitations. At the same time, their dependence on user-defined parameters, data choices, and method-specific assumptions creates significant scope for manipulation, ambiguity, and misunderstanding.

Beyond these technical and methodological variabilities, the reliance on historical training or background data, which is often essential for methods like SHAP, raises fundamental legal questions regarding the GDPR's purpose limitation principle (Art. 5(1)(b) GDPR) and the assessment of purpose compatibility (Art. 6(4) GDPR). Controllers bear the burden of proving that repurposing personal training data for post-hoc explanations is compatible with the original purpose of training the AI system; otherwise, such processing is prohibited. This poses a significant challenge, particularly when viewed against the backdrop of recent regulatory guidance: For example, guidelines from the German Data Protection Conference (DSK) emphasized that processing purposes must be strictly defined prior to use, creating risks for post-hoc explanation methods introduced after data collection (DSK, 2024). Furthermore, the European Data Protection Board (EDPB) distinguishes between the development and deployment phases of AI, warning that data usage across these phases constitutes separate processing activities that generally require distinct legal justifications (EDPB, 2024). This regulatory stance creates a "compliance trap": utilizing historical training data (development phase) to generate explanations during the process of application (deployment phase) risks violating the purpose limitation principle.

However, emerging case law suggests a pragmatic path that arguably points towards resolving this tension. A recent judgment by the Regional Court of Bayreuth ordered a credit bureau to disclose “how and with what weighting each datum entered into the score calculation”, specifically mandating a disclosure of “what the score would have been without consideration of [each specific] datum” (Bayreuth, 2025). While this specifically frames a counterfactual question (“Leave-One-Out”), it legally compels the controller to calculate the precise marginal impact of individual data points. By mandating such specific analyses to fulfill the right of access (Art. 15(1)(h) GDPR), the court adopts a pragmatic stance that presupposes the necessity of such calculations. Notably, the court, similar to the CJEU in *Dun & Bradstreet*, did not explicitly engage with the conflict regarding purpose limitation. Yet, by treating the generation of these explanations as an integral part of the transparency obligation, the rulings effectively presuppose purpose compatibility.

Against this complex background of technical instability and legal uncertainty, the weaknesses of current explanation methods create misaligned incentives between those who provide explanations and those who receive them (Martens et al., 2025). For credit institutions in their interactions with customers, explanations are not only instruments of transparency but also strategic tools that affect reputation, regulatory scrutiny, and customer responses. When the choice of explanatory method or background data influences how fair or comprehensible a model appears, lenders may have an incentive to select or frame explanations opportunistically. They may emphasize factors that appear controllable for consumers, such as higher income or lower credit utilization, while downplaying model or data-driven biases. For credit scoring companies, incentives are shaped primarily by their contractual relationships with banks rather than by direct accountability to consumers. In supervisory interactions, banks face yet another set of incentives, which may lead them to frame explanations in ways that minimize apparent model weaknesses (see on the related topic on internal credit risk models Begley et al., 2017). This can shift responsibility away from institutional design choices and towards individual borrowers or individual institutions, reducing pressure to improve models structurally. Producing detailed and legally robust explanations is also costly, especially for complex or proprietary systems, which can encourage under investment in genuine interpretability and reliance on explanations that formally satisfy transparency requirements without providing substantial insight.

These dynamics are likely to vary across institutions. Large incumbent banks with extensive compliance structures and in-house data science capabilities are often better positioned to generate explanations that are both compliant and reasonably intelligible. Smaller lenders and fintech firms that depend on third-party scoring providers may find it more difficult to meet the same standards. Differences in the capacity to deliver high-quality explanations can therefore translate into competitive advantages,

enabling well-resourced institutions to signal fairness and reliability more convincingly than less resourced competitors. Over time, consumers may self-select into institutions that offer clearer and more trustworthy explanations, which can deepen market segmentation (for the mechanism see Heidhues et al., 2017). This dynamic may inadvertently benefit traditional consumers who are in a position to compare offers and respond to differences in interest rates and disclosure quality. For more vulnerable groups such as refugees, who may face legal, linguistic, or documentation barriers, the relevant question is often not which institution to choose but whether they receive any formal offer at all, which can increase the risk of exclusion from mainstream credit or reliance on informal and high cost lenders where explainability is weakest or absent, thereby reinforcing informational inequality within the credit ecosystem.

In a similar vein, explanation mechanisms, similar to other transparency mechanisms, may also generate selection and behavioral distortions (see, e.g., Bauer & Gill, 2024). Individuals with higher financial literacy or better digital access may be more capable of interpreting explanations and adjusting their behavior strategically, for example, by optimizing visible variables while leaving underlying risk factors unchanged. Others who find explanations difficult to understand may react sub-optimally or withdraw from credit offers altogether. This unequal ability to act on explanations can unintentionally exacerbate existing disparities, as better-informed borrowers adapt while more vulnerable ones remain exposed. Without governance structures and trusted intermediaries that align incentives and standardize explanatory practices, explainability risks reproducing some of the information and power asymmetries it is intended to mitigate.

V Legal limitations

The legal rules that deal with credit scoring and underwriting, AI and data protection have so far not provided details on what might count as a good enough explanation of “the underlying logic” of a decision or the “role of the AI system” in decision-making procedures. Arguably, this has to do with legal rules being conceptualized against the background of human actors. They respond to humans, typically trying to understand why a certain outcome was produced instead of another, e.g., why someone received a low instead of a high credit score. More concretely, reference-dependence of, for instance, contrastive models allows for counterfactual reasoning where humans can ask whether an outcome would have been different had specific characteristics been different. This ties to notions of causal relationships between inputs and outputs. Psychologists have repeatedly stressed that “causal knowledge is used as the basis of predictions (...), action planning, decision-making, and problem solving”, hence, “explanation and causation are intimately related”. Given that “causal claims are often answers to (...) questions about why or how something occurred”, they have the potential to remedy what might otherwise seem as random

or unfair. From this perspective, providing causal explanations, at least when it comes to explaining individual predictions, can help acceptance of the outcome, allow for counterarguments, rectification, or ultimately some form of redress.

The problem with this legal approach is that a black-box will not produce causality along traditional lines. Both local and global explanation methods provide second-hand explanations, as it were, that construct a new model to explain the unexplainable black-box model. This is not to deny that the law might ultimately accept these methods as sufficient explanation. However, for now there is a lack of transparency about how explainable AI methods work, including the potential flaws in explanatory and causal power. Highlighting this will be a first and crucial element in socio-technical-legal discourse.

V.1 An XAI intermediary?

In the preceding sections we have identified technical, economic, and legal limitations of relying solely on feature attribution and counterfactual explanation methods. These limitations become particularly pronounced when stakeholders have divergent interests and when information asymmetries shape how explanations are produced, framed, and understood. One way to address these issues and to increase the likelihood that technical solutions deliver on their intended purposes could be to establish a trusted intermediary institution. In this last section of our essay, we outline thoughts on the nature and characteristics of such an intermediary

An XAI intermediary would securely hold model details and other required technical inputs to generate explanation reports for relevant stakeholders. In this arrangement, a neutral body would securely store the AI model, the necessary training data, and the accompanying documentation and would act as a bridge between model providers, affected individuals, and regulators. For example, a bank could share its credit scoring model and supporting materials on a confidential basis. Applicants could then request plain language explanations of specific decisions, while regulators and auditors would receive global reports that summarize model logic and highlight the most influential features. The intermediary would require an auditing authority and full access to the AI system's inner workings. It would generate explanations without disclosing the raw model or source code to the requester, thus protecting trade secrets while ensuring transparency.

Dun & Bradstreet further illustrates why an institutional intermediary would be necessary in practice. While the ECJ acknowledged the tension between access rights and trade secrecy, its proposed solution relies primarily on traditional judicial mechanisms. The judgment envisages an *in camera* procedure, stating that the “controller is required to provide the allegedly protected information to the competent supervisory authority or court, which must balance the rights and interests at issue with a view to

determining the extent of the data subject’s right of access provided for in Article 15 of the GDPR” (para. 76). This means that the controller must provide the contested information to a neutral authority. That authority, in turn, safeguards trade secrets while ensuring that all information falling within the scope of Article 15 is disclosed to the data subject, including details on the “logic involved”, which form part of the explanation obligation.

However, enforcement of Article 15 rights varies significantly across EU Member States. In Austria, where Article 15 can be enforced through both civil and administrative proceedings, *in camera* review by data protection authorities is possible. For example, Germany, on the other hand, restricts enforcement exclusively to civil courts where trade secret protection is possible only through § 273a ZPO confidentiality orders, not via *in camera* proceedings. These limited protections apply solely against third parties, not the opposing litigant, risking both excessive disclosure and uncertainty about whether explanations satisfy Article 15 requirements.

Such procedural constraints drive companies to defensive strategies. The German credit scoring company Schufa – which was involved in the predecessor case to Dun & Bradstreet – has adopted a pragmatic approach: instead of risking disclosure of sophisticated AI models through civil litigation, it relies solely on logistic regression and gradient tree boosting (Schufa, n.d.), offering consumers simplified, app-based score explanations while disclosing full details only to data protection authorities.

Apart from this procedural fragmentation, requiring courts and supervisory authorities to determine “the extent of the data subject’s right of access” requires a level of technical understanding that courts and authorities frequently lack. In the context of algorithmic scoring, defining the scope of necessary disclosure is not merely a legal balancing act but a technical assessment of what constitutes a meaningful explanation. These institutions typically lack the deep expertise in explainable AI methodologies required to decide whether a particular set of feature attributions or counterfactuals provides a sufficient understanding of the “logic involved”.

Along these lines, we propose to design an intermediary with which the trained models and used training data would need to be shared. This intermediary would then decide what information must be shared with individuals. Acting as an expert authority, the XAI intermediary could operationalize the balancing test, protecting trade secrets while ensuring that the data subject’s right to information under Article 15 GDPR is meaningfully fulfilled, thereby standardizing the balance between explanation and intellectual property protection.

Designing such an XAI intermediary requires mechanisms to ensure independence and trustworthiness. Fiduciary duties would uphold data subjects’ rights, prioritizing their interests, keeping data minimization

and risks of bias in mind, while at the same time respecting the legitimate interests of model owners by not disclosing data and model details to the broad public. If **explanations are treated as a public good**, the intermediary could be a public agency, ensuring high independence and legal authority to compel information, thereby fostering public trust. At the same time, public agencies may face limitations in capacity and expertise, and their processes can be slow or inflexible.

From a technical perspective, the intermediary should operate secure data centers, maintaining redundant copies of all information to ensure service continuity. Models and training data should be stored in isolated, highly protected environments accessible only to a vetted subset of staff, with sensitive computations performed in secure enclaves to prevent unauthorized access. The platform would require sufficient computing resources, including specialized processors for intensive analysis, and long-term, write-once storage for immutable audit trails. Data should flow through controlled channels, including secure connections to client systems, comprehensive change logs, and catalogs of datasets, models, and reports. Personal data collection should be minimized, names should be tokenized where possible, and synthetic data should be used for testing. Strong encryption must protect data at rest and in transit, with encryption keys secured in dedicated hardware. Access should be tightly controlled with verification of every request, temporary permissions, multi-factor authentication, continuous security monitoring, proactive attack protection, regular backups, and tamper-evident logs. For accessibility, the platform should provide two web portals: a public, multilingual interface for identity verification, request submission, status tracking, and receipt of explanations for data subjects; and a regulated partner dashboard for model uploads and audit management for supervisors. To achieve scalability, the intermediary should leverage automated XAI tools and processes. However, full automation risks overlooking context or nuance. Hence, a human-in-the-loop approach is essential, where experts such as data scientists, XAI specialists, and ethicists oversee the automated analysis, particularly for complex or borderline cases. These experts can verify that the generated explanations are meaningful and not misleading.

With this infrastructure, the intermediary can deliver “explainability reports” at two levels. For individuals affected by AI-based decisions, the intermediary provides a local explanation report focused on the reasoning behind a specific prediction. This report could combine feature attribution (e.g., SHAP values showing which features contributed most to a decision) and counterfactuals (concrete scenarios showing what changes would alter the outcome). For example, a report might explain that a credit application was denied due to low income and past delinquencies, and indicate that saving an additional \$1,000 would likely have resulted in approval. These reports can present key features and counterfactual examples in accessible formats. At the same time, the explainability report would need to make it transparent that it

is delivering a second-hand explanation without access to the original model's black-box. Consumers would need to understand what they can learn from the report and how likely it is that behavioral change would actually change their score.

For regulators, auditors, or oversight bodies, the intermediary could produce global explanation reports that summarize the AI system's overall logic and behavior. Such reports are vital for compliance with regulations such as the current version of the EU AI Act. The global report can include a model overview, an analysis of the most influential input features, performance metrics broken down by demographic group, and evidence supporting compliance, all without exposing proprietary details. This approach helps regulators assess systems without forcing companies to reveal trade secrets.

To demonstrate effectiveness, pilot programs with select providers and authorities can validate demand, refine pricing, and calibrate service levels. Key success indicators include report turnaround time, rates of resolving contested decisions, findings of noncompliance, and user satisfaction. Through this approach, the intermediary can turn explainability into a standard service that safeguards rights, promotes innovation, and reduces regulatory risk.

V.2 Concluding thoughts on Explainability-as-a-Service

Alternatively, the intermediary could operate as a business, offering "Explainability-as-a-Service". Companies would pay for analysis and certified explanation reports. As regulation tightens, demonstrating external auditing and explainability could become a market differentiator, motivating adoption even without strict legal requirements. To prevent conflicts of interest, particularly when a for-profit intermediary is paid by the model owner, safeguards such as regulatory oversight, audit standards, and rotation requirements are necessary to ensure objectivity and accuracy. The establishment of a private EaaS intermediary, however, raises legal concerns.

First, the use of a private EaaS intermediary could not absolve the controller of their obligation to provide access to information under Art. 15 GDPR by themselves. Since the data subject receives a report directly from the EaaS provider, the controller is prevented from selectively presenting data to ensure a compliant appearance. However, this dual flow of information creates a specific risk: If the mandatory explanation provided by the controller diverges from the independent EaaS report, the former may appear misleading or inadequate. This exposes the controller to considerable liability risks, as data subjects can leverage such discrepancies to claim damages or trigger administrative fines.

Second, the generation of meaningful model explanations often requires access not only to the deployed model but also to parts of the original training data, e.g., for SHAP, which frequently contain personal

data. While this access could, in theory, be facilitated via an API, such an approach entails significant security risks, and legal frameworks such as NIS2 and DORA may impose particularly stringent cybersecurity requirements. As a result, full data transfers of both the model and the training data to the intermediary may become necessary. If the training data includes personal data, this raises the question of whether the intermediary could qualify as a data processor, acting solely on behalf of and under the instructions of the deploying entity. This could reduce the internal compliance burden for the deploying company; however, processors are, by definition, bound to follow the instructions of the controller and lack discretion over the purposes and means of processing, which restricts the intermediary's ability to act as an independent evaluator. A more suitable legal framework might be that of joint controllership; however, this would also bring significantly increased regulatory obligations, particularly regarding lawful data transfers and accountability mechanisms.

References

- Bastos, J. A., & Matos, S. M. (2022). Explainable models of credit losses. *European Journal of Operational Research*, 301(1), 386-394.
- Bauer, K., Hinz, O., Van Der Aalst, W., & Weinhardt, C. (2021). Explaining AI to me—explainable AI and information systems research. *Business & Information Systems Engineering*, 63(2), 79.
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Explaining AI: The impact of explainable artificial intelligence on users' information processing. *Information systems research*, 34(4), 1582-1602.
- Bauer, K., & Gill, A. (2024). Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. *Information Systems Research*, 35(1), 226-248.
- Begley, T., Purnanandam, A., & Zheng, K. (2017). The Strategic Underreporting of Bank Risk. *The Review of Financial Studies*, Volume 30, Issue 10, 3376–3415.
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7), 2845-2897.
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89(1), 1–34.
- Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020, August). Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature* (pp. 448-469). Cham: Springer International Publishing.
- De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management*, 15(12), 556.
- Dobbie, W., Goldsmith-Pinkham, P., Mahoney, N., & Song, J. (2020). Bad credit, no problem? Credit and labor market consequences of bad credit reports. *The Journal of Finance* 75.5, 2377-2419.
- Dobbie, W., & Skiba, P. (2013). Information Asymmetries in Consumer Credit Markets: Evidence from Payday Lending. *American Economic Journal: Applied Economics* 5 (4), 256–82.

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

Dun & Bradstreet Austria, Case C-203/22 (Court of Justice of the European Union February 27, 2025). <https://curia.europa.eu/juris/document/document.jsf?text=&docid=295841&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1>

European Central Bank. (2025). ECB guide to internal models (Release 4.0). https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining data-driven decisions made by AI systems: The counterfactual approach. *MIS Quarterly*, 46(3), 1635-1660.

Frost, J., Gambacorta, L., Huang, Y., Song Shin, H., & Zbinden, P. (2019). BigTech and the changing structure of financial intermediation, *Economic Policy*, Volume 34, Issue 100, 761–799.

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

Gibbs, C., Guttman-Kenney, B., Lee, D., Nelson, S., Van der Klaauw, W., & Wang, J. (2025). Consumer credit reporting data. *Journal of economic literature*, 63(2), 598-636.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4, 752558.

Heidhues, P., Kőszegi, B., & Murooka, T. (2016). Inferior products and profitable deception. *The Review of Economic Studies*, 84(1), 323-356.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3), 388-409.

Lakkaraju, H., & Bastani, O. (2020, February). "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79-85).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., ... & Provost, F. (2025). Beware of "explanations" of AI. arXiv preprint arXiv:2504.06791.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information systems management*, 39(1), 53-63.

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

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607-617).

- Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., & Welke, P. (2023, September). An empirical evaluation of the Rashomon effect in explainable machine learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 462-478). Cham: Springer Nature Switzerland.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-52).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
- Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous agents and multi-agent systems*, 33(6), 673-705.
- 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.
- Sargeant, H. (2026). Mind the gap: Securing algorithmic explainability for credit decisions beyond the UK GDPR. *Computer Law & Security Review*, 60, 106247.
- SCHUFA Holding AG. (n.d.). Welche Verfahren nutzt die SCHUFA für das Bonitätsscoring? [Which procedures does SCHUFA use for credit scoring?]. Retrieved December 11, 2025, from <https://www.schufa.de/faq/privatpersonen/scoring/welche-verfahren-nutzt-die-schufa-fuer-das-bonitaetsscoring.jsp>
- Stiglitz, J. E., and A. Weiss. "Asymmetric Information in Credit Markets and Its Implications for Macroeconomics." *Oxford Economic Papers* 44, no. 4 (1992): 694–724.
- Szwabe, A., & Misiorek, P. (2018, August). Decision trees as interpretable bank credit scoring models. In International Conference: Beyond Databases, Architectures and Structures (pp. 207-219). Cham: Springer International Publishing.
- von Zahn, M., Liebich, L., Jussupow, E., Hinz, O. & Bauer, K. (2025). Knowing (Not) to Know: Explainable Artificial Intelligence and Human Metacognition. *Information Systems Research* 0(0).
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, 7(2), 76-99.
- Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, 10(2), e0117844.