

Kevin Bauer  
Lucia Franke  
Andrej Gill  
Katja Langenbucher

# Institutionalizing Explainability in Credit Scoring

SAFE Policy Letter No. 114 | May 2026

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

[policy\\_center@safe-frankfurt.de](mailto:policy_center@safe-frankfurt.de) | [www.safe-frankfurt.de](http://www.safe-frankfurt.de)

## Institutionalizing Explainability in Credit Scoring\*

Kevin Bauer<sup>†</sup> Lucia Franke<sup>‡</sup> Andrej Gill<sup>§</sup> Katja Langenbucher<sup>¶</sup>

May 6, 2026

### Summary

Machine learning credit scoring expands the informational frontier of retail lending, particularly for thin file borrowers, yet it also erodes the practical meaning of disclosure duties that anchor consumer protection and prudential oversight. The central financial implication is that explainability is no longer a peripheral communication task. It is a market structuring variable that can reshape access, pricing efficiency, competition, and the distribution of compliance burdens across incumbents and challengers. The central regulatory gap is that current regimes articulate rights and obligations but remain under specified on what constitutes a sufficient explanation, how fidelity can be verified, and how opportunistic framing can be prevented when explanation techniques permit multiple plausible narratives. The most effective policy response is institutional rather than purely technical, achieved by creating a governed intermediary layer that can translate proprietary model behavior into standardized consumer facing and supervisor facing disclosures while preserving legitimate confidentiality.

---

\*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.

<sup>†</sup>Goethe University Frankfurt (Professor for Game-Theoretic and Causal AI in Business and Economics), Leibniz Institute for Financial Research SAFE, Hessian.AI, Email: [bauer@safe-frankfurt.de](mailto:bauer@safe-frankfurt.de)

<sup>‡</sup>Goethe University Frankfurt, Email: [franke@jur.uni-frankfurt.de](mailto:franke@jur.uni-frankfurt.de)

<sup>§</sup>Gutenberg-University Mainz, Email: [gill@uni-mainz.de](mailto:gill@uni-mainz.de)

<sup>¶</sup>Goethe University Frankfurt (Professor of Civil Law and Financial Market Regulation), Leibniz Institute for Financial Research SAFE, Institute for Monetary and Financial Stability (IMFS), Email: [langenbucher@safe-frankfurt.de](mailto:langenbucher@safe-frankfurt.de)

# 1 Context and Problem

Retail credit markets have always depended on information processing to manage adverse selection and moral hazard. The shift in the last decade is the scale, granularity, and opacity of the signals that lenders can lawfully incorporate. High dimensional behavioral and transactional data combined with flexible machine learning models can improve risk discrimination and expand credit access for populations that lack conventional signals such as long bureau histories, stable local employment, or locally legible documentation. That promise, however, is accompanied by a governance deficit. When underwriting is driven by complex models whose internal logic is difficult to reconstruct in human terms, transparency obligations can collapse into formalistic disclosures that do not enable contestation, error correction, or credible supervisory review.

This policy letter translates the core insights from the research study: “Institutionalizing Explainability: On Credit Scoring AI and Consumer Agency” by Kevin Bauer, Lucia Franke, Andrej Gill and Katja Langenbucher into a regulatory agenda for financial authorities and credit scoring experts who must reconcile machine learning driven underwriting with enforceable transparency, consumer agency, and supervisory control.

Regulators face a triangulated challenge that is sharper in credit than in many other algorithmic domains. Consumers possess actionable rights to meaningful information about adverse decisions and the logic involved in automated processing, while firms retain legitimate interests in protecting trade secrets and preventing gaming of underwriting systems. Supervisors, meanwhile, require a systemic view that individual adverse action notices cannot provide, particularly when discrimination risks arise through proxy variables and non linear interactions rather than explicit use of protected attributes. Courts and lawmakers are increasingly unwilling to allow trade secrecy to nullify informational rights, yet legal systems are often not equipped to adjudicate what constitutes an intelligible explanation for a black box model, nor to do so at the scale of retail credit markets.

This tension has direct consequences for market structure and model choice. When disclosure duties are unpredictable and procedural protections for confidentiality vary across jurisdictions, firms rationally pursue defensive strategies. These include simplifying models to reduce explanation exposure, restricting the use of certain techniques, and narrowing underwriting criteria to avoid contested edge cases. The result can be mispricing, reduced inclusion, and a competitive tilt toward institutions with the compliance infrastructure to operationalize explainability at scale. In other words, explainability is becoming a channel through which regulatory design choices influence both innovation incentives and distributional outcomes in credit access.

## 2 Analysis

The key analytical point for regulators is that explainability in credit scoring is socio technical rather than purely technical. Explanations are generated, presented, and interpreted by actors with divergent incentives. Developers tend to use explainability tools for debugging and plausibility checks, lenders face compliance and reputational exposure, consumers need actionable information for contestation and behavioral adaptation, and supervisors require aggregate insight for systemic oversight. A single explanation format cannot satisfy these heterogeneous

demands, and attempts to compress them into a one size disclosure can produce information overload for consumers while still leaving supervisors without a reliable signal of model behavior. The most widely deployed explainability approaches in modern credit scoring are post hoc, meaning they are layered on top of an already trained model rather than being intrinsic to a transparent model form. Feature attribution methods decompose a prediction into estimated contributions of input variables, while counterfactual methods propose changes to inputs that would alter an outcome. These approaches can be useful, but they must be understood as approximations produced by auxiliary procedures rather than direct readings of model logic. Their outputs depend on parameter choices, reference distributions, and representation decisions that recipients rarely see. This dependency generates an explanation integrity problem that is particularly acute in credit.

Integrity fails in two ways. The first is instability. Different explainability methods, or the same method with different background datasets, can yield materially different accounts of why a decision was made. That variability undermines the consumer's ability to contest an outcome and the lender's ability to treat explanations as reliable diagnostics within model risk management. The second is discretion. Where multiple plausible explanations exist, institutions can select narratives that shift responsibility onto borrowers, emphasize controllable factors that are difficult or impossible to change in practice, or mask proxy discrimination by choosing methodological settings that dampen the apparent influence of sensitive attributes. Even when there is no intent to mislead, these degrees of freedom create a structural risk that explanations become rhetorically compliant artifacts rather than governance instruments.

Legal doctrine compounds these technical limitations. Producing meaningful explanations can require access to training data or reference datasets that contain personal data, raising purpose limitation and phase separation concerns under data protection law. The practical reality is that transparency duties increasingly presuppose marginal contribution style analysis, yet the lawful basis for repurposing historical data for explanation may be contested unless clarified by regulators. This creates a compliance trap in which robust explanation is simultaneously demanded and procedurally discouraged, incentivizing minimal disclosure, model simplification, or selective framing rather than substantive accountability. These dynamics point to a deeper institutional conclusion. The credit ecosystem needs a trusted mechanism that can absorb information sensitive complexity, reduce duplication of verification costs, and translate proprietary model behavior into decision relevant signals for different audiences. Classic theories of financial intermediation describe delegated monitoring as a solution to information asymmetry and verification cost problems. The same logic applies to algorithmic underwriting, where individual consumers cannot feasibly verify model behavior and where supervisors cannot scale bespoke judicial balancing of transparency and secrecy. An intermediary layer can also play a gatekeeping role by staking institutional capital on the integrity of the explanation process, but that role cannot rely on reputation alone because quality is hard to observe and because mandated certification can degrade into a mere regulatory license when comparability and testing standards are weak.

### 3 Policy Recommendations

Regulators should treat explainability as regulated informational infrastructure for credit markets rather than as discretionary customer communication. A central recommendation is the creation of an ex ante accreditation regime for explainability intermediaries that front loads competence, independence, cybersecurity, and quality management. This approach is superior to reliance on ex post liability because courts are unlikely to have stable technical yardsticks for judging explanation quality, and because the expectation gap between what a certification implies to the public and what it actually tests is wider in the presence of black box uncertainty. Accreditation should prohibit conflicts that collapse independence, especially the combination of model design consulting and explanation certification for the same client within the same organizational boundary. It should also incorporate enhanced scrutiny where an intermediary becomes financially dependent on a small set of large clients, since fee concentration predictably weakens the incentive to issue unfavorable findings.

A second recommendation is to mandate a standardized two audience reporting architecture that separates individual contestation focused disclosure from supervisory system level disclosure, while making methodological choices observable and auditable. Consumer facing reports should provide intelligible local explanations that combine feature contribution information with realistic counterfactual pathways, accompanied by clear statements of uncertainty and limits so that consumers do not mistake post hoc approximation for causal proof. Supervisor facing reports should provide global insight into model behavior, including aggregate feature reliance patterns, stability diagnostics, and fairness metrics that allow oversight of disparate impact risks without exposing proprietary source code. Standardization reduces strategic selectivity, improves cross firm comparability, and enables benchmarking. It also solves the observability problem that undermines reputational discipline by making intermediary performance legible to supervisors and, where appropriate, to the public in aggregated form.

A third recommendation is to implement secure access and data governance requirements that enable meaningful explainability without forcing broad disclosure of trade secrets and without allowing explanation to become a pretext for uncontrolled secondary use of data. Intermediaries should operate within secure data vault architectures that use strict access control, tamper evident audit logs, encryption in transit and at rest, tokenization where feasible, and tightly scoped computational environments for sensitive operations. These operational requirements should be harmonized with financial sector cyber resilience expectations so that explanation generation is treated as a core risk function. In parallel, regulators should clarify the lawful basis and purpose compatibility conditions under which training data and model artifacts may be processed for explanation, thereby preventing a situation where lawful explanation is technically infeasible and institutions respond by retreating from inclusive modeling.

These institutional levers must be integrated into supervisory practice. Firms should be required to operationalize explanation outputs within complaint handling, adverse action workflows, internal override governance, and model risk management so that explainability becomes part of a closed loop control system. The intermediary layer does not displace accountability of deployers. It restructures the informational pathway through which accountability is evidenced, tested, and improved.

## 4 Conclusion

Machine learning credit scoring is pushing credit markets toward a new equilibrium in which high dimensional prediction and enforceable transparency are simultaneously demanded, yet remain structurally in tension. Post hoc explainability techniques can support contestation and consumer learning, but their instability, discretion, and legal entanglements make them unreliable as the sole foundation for accountability. The regulatory task is therefore to institutionalize explainability so that it becomes a credible governance capability rather than a performative disclosure.

A regulated intermediary regime anchored in ex ante accreditation, standardized two audience reporting, and secure data governance offers a coherent path forward. It can preserve innovation compatible underwriting while protecting consumer agency, limiting opportunistic framing, and strengthening supervisory visibility. For regulators and credit scoring experts, the decisive question is not whether explainability will be required, since legal pressure is already moving decisively in that direction, but whether it will be organized in a way that sustains inclusion, competition, and legitimacy in the credit economy.