Explainable AI as a Component of Building Trust

The Case of Regulating Creditscoring

Katja Langenbucher

Abstract *The paper takes up the notions of trust and explainability in the GDPR and in upcoming German legislation, using AI-based credit scoring as an illustration. It offers an overview of methods of explainable AI, stressing differences between computer scientists, legal scholars, and legislators. Counterfactual explainability, the paper claims, might be useful along the lines of the ECJ decision (Court of Justice of the European Union 2025).*

> "The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development [...] and the use of artificial intelligence systems (AI systems) in the Union in accordance with Union values to promote the uptake of human centric and trustworthy artificial intelligence". This is how the first recital of EU Regulation 2024/1689 of 13 June 2024 (AI Act) on harmonized rules concerning artificial intelligence (AI) starts. At what point an AI system counts as trustworthy is not defined in the AI Act. Instead, the term appears in the law in a variety of contexts. We find, for instance, the uptake of "human centric and trustworthy" AI (Recital 1, AI Act), the goal to develop "secure, trustworthy and ethical AI" (Recital 8, AI Act), along with "accuracy, reliability and transparency [...] to avoid adverse impacts, retain public trust and ensure accountability and effective redress" (Recital 59, AI Act).

1. The Concept of Trust in the AI Act

What counts as "trustworthy" varies significantly across disciplines and context (see Kaminski 2025 on philosophy; Zhang et al. 2024 on psychology; Aljohani et al. 2025 on medicine; Breuer and McDermott 2008 on economics). The AI Act does not define the concept of trust or of trustworthiness. Instead, it mostly appears as an element of explaining the EU Commission's regulatory philosophy, based on everyday language.

Trust. Arguably, one of the first times the term "trust" surfaces in the context of AI regulation is in the 2018 EU Commission Communication "Artificial Intelligence for Europe" (EU Commission 2018). That strategy references the GDPR as a "*major step for building trust, essential in the long-term for both, people and companies*", along with the – then – proposals for the flow of non-personal data, the e-Privacy Regulation and the Cybersecurity Act. Additionally, the Communication emphasizes the role of private rights of actions if things go wrong: "*A high level of safety and an efficient redress mechanism for victims in case of damages helps to build user trust and social acceptance of these technologies*".

Trustworthiness. A year later, the 2019 High Level Expert Group Ethics Guidelines for Trustworthy AI, endorsed by the EU Commission (EU Commission 2019), introduced "*trustworthy AI*" as a guiding concept. Trustworthy AI "*should be (1) lawful – respecting all applicable laws and regulations, (2) ethical – respecting ethical principles and values, (3) robust – both from a technical perspective while taking into account its social environment*".

An ecosystem of trust. In its 2020 White Paper on AI (EU Commission 2020), the Commission further developed the concept into one of two pillars of its AI Regulation. The first one is an "*ecosystem of excellence*", the second an "*ecosystem of trust*". The latter is "*a policy objective in itself*" and "*should give citizens the confidence to take up AI applications and give companies and public organizations the legal certainty to innovate using AI*". Along those lines, the AI Act uses the term as a goal, justifying the Act's risk-based approach: "*To ensure a high level of trustworthiness, certain mandatory requirements should apply to high-risk AI systems*" (Recital 64, AI Act). "*While the risk-based approach is the basis for a proportionate and effective set of binding rules*", Recital 27 stipulates, "*it is important to recall the 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG appointed by the Commission. In those guidelines the AI HLEG developed seven non-binding ethical principles for AI […]. These seven principles include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability*".

Transparency. One of the seven principles of trustworthy AI is transparency. "Transparency", we find in that same Recital 27, "means that AI systems are developed and used in a way that allows appropriate traceability and explainability". Traceability targets compliance, documentation and performance during the lifetime of an AI system (Recitals 27, 53, 71, Art. 12(2)). Explainability, by contrast, references – like in 2018 – the link between trust and private rights of action. The "exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defense and the presumption of innocence" requires the citizen to have appropriate information to back up a litigation claim. Faced with AI, this can be challenging, if the potential litigant has no understanding of what triggered a particular decision and who might be responsible for it.

Explainability. Against this background, explainability is a core component of trust-based AI regulation. This is not to be understood as a *necessary* element. Arguably, in most low-risk use cases, consumers do not care about an explanation of how the AI produced its result. Picture-generating models provide an illustration: Using an AI to design a birthday card or enhance a power point presentation does not call for an explanation of how the AI did it, as long as the user enjoys the picture. This is different for high-risk use cases. A person who receives a lower credit score than he or she expected is likely to demand explainability of what led to the score, to change behavior or to prepare litigation.

2. Explainability Rights

Art. 22, 15 GDPR provide core private rights, with the AI Act and (for credit-scoring) the Consumer Credit Directive adding finishing touches, as it were. Art. 22 GDPR regulates instances of automated decision-making that produce "*legal effects* […] *or similarly significantly*" affect the data subject. In line with the Regulation's general approach, the rule starts with a prohibition of this type of automated decision-making, Art. 22(1) GDPR. Then, Art. 22(2), (3) GDPR follow up with exceptions to the ground rule.

Explainability of automated decision-making is covered in Art. 15(1)(h) GDPR. The rule provides the data subject with a right to "*meaningful information about the logic involved*". While the legislator might not have had AI-based decision-making in mind, the text of the Regulation covers it, and two recent ECJ-decisions, both concerning credit-scoring, have broadened, rather than narrowed, its scope.

Against this background, it is unsurprising that it was late in the process of passing the AI Act, that EU legislators decided to, in addition, enshrine a right to an explanation of individual decision-making in Art. 86(1): "*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* […] *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*".

The text seems to mostly repeat Art. 15 GDPR. Both rules concern rights to receive an explanation, and Art. 86(3) AI Act mentions that its para. (1) shall not apply if the right it confers is otherwise provided for under Union law. Automated decision-making, including profiling, triggers Art. 15(1)(h) GDPR. Rather vaguely, it then speaks about access to information about *the logic involved*. Not each instance of automated decision-making involves an "AI system" under Art. 3(1) AI Act, which requires "*varying levels of autonomy*" and inferences from the input the system receives on *how to generate outputs such as predictions, content, recommendations, or decisions*. In that sense, the scope of Art. 86(1) AI Act is narrower, because it applies only to AI systems. There is a slight variation in the text, if compared to Art. 15(1)(h) GDPR: The AI Act concerns a right to obtain an explanation on "*the role of the AI system in the decision-making procedure*" (*sur le role du système d'IA dans la procedure décisionelle; zur Rolle des KI-Systems im Entscheidungsprozess; sul ruolo del sistema di IA nella procedura decisionale)*. The GDPR, by contrast, focuses on the *logic involved* in the automated decision-making, including profiling. One might read this as the AI Act being more interested in an understanding of what the AI system contributes to the overall decision-making process, whereas the GDPR has the functioning of the AI itself in mind. However, the wording is similar and, arguably, Art. 86(1) AI Act will mostly be relevant to fill gaps Art. 15(1)(h) GDPR might leave.

3. An illustration: AI-based Credit Scoring

Like the concept of "trust", the concept of "explainability" varies according to context. This makes an inductive approach useful, presenting one example for the role of explainability in the general context of trustworthy AI. AI-based credit scoring is an apt illustration: It qualifies as a high-risk AI-system under Art. 6(2), Annex III No. 5 AI Act. Building a credit score involves large amounts of data, which brings the GDPR into play. Two recent European Court of Justice decisions have found that computing a credit score counts as automated decision-making under Art. 22 GDPR, hence, Art. 15(1)(h) GDPR's right to be informed about the "logic involved" applies. Additionally, there is sectoral EU legislation concerning creditors and German legislation in the making on credit scoring.

Which role does explainability play in the context of credit scoring? A consumer might, for several reasons, ask for an "explanation": To verify the accuracy of the data used, to decide whether he has given his consent for data use, to adapt his behavior, in the hope of receiving a better score in the future, or to litigate, should the score seem unfairly low. In that way, an "explainable" credit score can be conducive to building trust.

The Consumer Credit Directive. Directive (EU) 2023/2225 (CCD) provides a sectoral private right of action, complementing the GDPR. It is different from the GDPR in that its scope extends solely to the relationship between creditor and borrower. Scoring agencies fall outside the CCD. It starts from the assumption that "*artificial intelligence (AI) systems can be easily deployed in multiple sectors of the economy and society*" (Recital 46, CCD). Following up on the GDPR, the CCD 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*" (Recital 56, CCD). Art. 18(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".

The German Bundesdatenschutzgesetz-draft. In Germany, legislators have been considering private rights of action under national law to further support the consumer in a credit underwriting situation. To this end, a law on credit scoring (drafted before the new government started in 2025, BT-Drucksache 20/10859) takes advantage of the discretion that the GDPR leaves for Member States. § 37a *Bundesdatenschutzgesetz*-draft, first, references Art. 22 GDPR which allows automated decision-making, such as credit scoring, if Member State law lays down the details. Second, § 37a *Bundesdatenschutzgesetz*-draft requires an "input control". It specifies certain data points, for instance, sensitive data under Art. 9 GDPR, that may not be used by a credit scoring company. The question whether an input control is a smart form of legislating is beyond the scope of this paper. Suffice to say that, given inferences that can be drawn from other data points (the problem of redundant encoding, see Barocas and Selbst 2016), this strategy will only help if an AI system works with a very limited set of data points (see Solove 2024 for the argument, that the concept of sensitive data is at a dead end). Third, the German draft law stipulates a form of quality control. The data used must be processed "*on the basis of an appropriate mathematical-statistical procedure which is demonstrably relevant to compute the probability of a specific conduct*".

Fourth and last, the rule includes its own version of a right to transparency. The company, which produces a credit score, must deliver information "*in a* precise, *transparent, understandable and easily accessible form as well as in clear and simple language*". Four elements must be explained: "*the personal data used, the weight of data points that influence the score most importantly, the meaning of the specific score and the score itself*". The law, which has not yet entered into force and may be changed under the new government, includes detailed explanations of what it has in mind as to transparency rights. More specifically, it requires scoring companies to use language that is targeted to its audience and to reflect upon its "cognitive capabilities".

4. "Explaining AI" – A short overview

Providing an explanation of automated decision-making is clearly at the forefront of the legislative endeavors mentioned in the previous paragraph. This (mostly correctly) presumes that those affected from automated decision-making have a general interest to understand how the decision was made (but see Langenbucher 2024 on "black-box rights"). However, the details of a transparency-enhancing private right of action depend very much on context. Understanding why a doctor suggests staying calm, after his AI predicted a low probability that a beauty mark is cancerous, raises very different issues than following up on an AI-based credit score or evaluating an AI-supported judicial decision to let someone go free on bail. In a low-risk environment, understanding an AI's inner workings might not be relevant and might not justify the drop in predictive performance that is often associated with explainable AI (Molnar 2022: 3.1.). By contrast, the necessities of safety measures and testing, the detection of bias or the wish to increase social acceptance might call for the use of explainable systems (ibid.: 3.1.).

Against this background, it is tempting to draw on computer science efforts to provide "explainable" AI (XAI). However, it is important to bear in mind that, as stressed above, context matters. A computer scientist who employs an XAI model will be interested in different questions than a consumer looking at his score, a lawyer preparing an anti-discrimination lawsuit, or a banking supervisor who runs a risk-management check. Some of this has to do with varying competences and capabilities of the actors (Kaminski 2025). Additionally, their specific reasons for requiring an explanation determine what is useful for them. The computer scientist will wish to gain a better understanding of the steps the AI takes, for instance, across the different layers of a neural network. Whether the outcome adequately represents an individual's capability to pay back a loan is not the computer scientist's concern, especially, if possible flaws have nothing to do with the model, but go back to faulty data. By contrast, neither the consumer nor the lawyer are overly interested in the inner workings of the model. Their core interest will usually lie with the data used and the predictive power of the score. The banking supervisor's interest is situated somewhere in between. Data that produce inappropriate uncertainty as to adequate representation of a portfolio of creditors may not be used. The same goes for models that are inappropriate for that purpose.

Explainability is not the same as substantive control. I might very well understand how a decision was made but still consider it unfair or unlawful. Often, explainability tells us something about the procedure of decision-making. This might be done, for instance, by reproducing each step, by highlighting core elements, or by producing counterfactuals. When discussing explanations in the context of AI, it is helpful to distinguish between two approaches: Using models that are inherently interpretable due to their simpler structure (like linear regression, decision trees, or k-nearest neighbors) and using post-hoc explanatory techniques designed to shed light on the behavior of more complex, often opaque black-box models (like deep neural networks). The field of XAI is primarily concerned with the latter, developing methods like LIME, SHAP, or DiCE (see Dubovitskaya and Bosold 2025). Efforts of credit scoring companies, such as the German SCHUFA, rely on the former in explaining consumers the impact of individual components of their credit score.

What all these situations have in common is the legislator's assumption that the contribution of each feature a model uses can be computed and disclosed. For linear regression models, this is correct (Molnar 2022: 9.5.1.). For more complex, nonlinear models, statisticians have produced a wide range of potential explanations that vary in usefulness according to context (overview at Freiesleben 2022; pointing to the irony that XAI models add a second layer of complex models on the blackbox model: Nisevic, Cuypers and De Bruyne 2025 noting an "XAI chaos").

Local surrogate methods. One common approach of XAI are local surrogate methods (Dubovitskaya and Bosold 2025). These methods produce a local explanation of the AI decision. An explanation is local, rather than global, if it targets the region around the prediction of interest. What it produces does not attempt to explain the inner workings of the entire model, but only one specific prediction, in our case: the credit score for a specific customer.

These methods are called surrogate, because they involve creating a simpler, interpretable model that approximates the behavior of a more complex model in the local region. The first step is to create a set of perturbed samples around the prediction point. This includes varying all features randomly within a local neighborhood. The samples are weighted according to their distance from the original point, with closer samples receiving higher weights to ensure that the surrogate model focuses on the immediate vicinity of the prediction. For these samples, the original, complex model is run to predict outcomes. Then, one trains a simpler, inherently interpretable 'surrogate' model on these perturbed samples and their predictions, aiming to mimic the complex model's behavior locally. This produces a local, surrogate explanation of the black-box model. A common choice for this surrogate model is a decision tree because these are relatively easy to understand. One takes the set of perturbed samples and their corresponding predictions from the complex model and trains a simple decision tree model using this local data. This trained decision tree then serves as the local explanation.

Given that a surrogate model does not try to explain the computation of the complex model, it is model-agnostic. This means it can (locally) explain predictions from any complex model. Taking these characteristics together, the technique is often called LIME (Local Interpretable Model-agnostic Explanations; Ribeiro, Singh and Guestrin 2016).

Advantages of local surrogate models, apart from being model-agnostic, are that they are simple to understand and capture the model's behavior around the specific point of interest, which might be different from its global behavior. It takes a shot at highlighting feature importance, something that, arguably, § 37a *Bundesdatenschutzgesetz*-draft was looking for, when it asked for "weights". A decision-tree algorithm determines, for instance, which features are most informative for splitting the data to match the complex model's predictions. If changing a feature significantly alters the complex model's predictions, for instance, the number of credit cards or open bills, the decision tree is likely to use that feature in its top-level splits. To make sure that the feature importance isn't skewed, one would perturb all features uniformly, vary the perturbation strategy, and do multiple runs. Still, a model like LIME assumes linear behavior of the model locally. It is unclear whether there is a solid theoretical basis for this assumption (doubting this: Molnar 2022: 9.5.4.).

Shapley additive explanations. Another powerful method to explain a complex model's predictions is SHAP (SHapley Additive exPlanations). This method is based on cooperative game theory. SHAP assumes that each feature value is a player in a game where the prediction is the payout. Shapley values in cooperative games demonstrate how to fairly distribute that profit among all players. Used to explain an ML-prediction, the first step is to single out players: Each individual feature the model uses counts as one player. Second, the prediction becomes the game's "payout" (ibid.: 9.5.1.). In this way, SHAP calculates the contribution of each feature to the actual prediction that the model arrived at by systematically including and excluding features to simulate different scenarios.

SHAP starts with a baseline prediction: the average model output over the entire dataset, for instance, an average score. Then, SHAP calculates how each feature, such as the number of credit cards a consumer holds, pushes the prediction away from the baseline. To do this, one generates perturbed samples (see ibid.: 9.5.5.) where each sample represents a different combination of features being "present" or "absent". One then inputs each sample into the complex model and records each output prediction. These output predictions help to understand how the (original, complex) model behaves in the local neighborhood around the relevant situation. By comparing how the prediction changes as features are perturbed, SHAP can deduce the importance of each individual feature. The predictions for the perturbed samples are compared against the prediction for the original sample and the marginal contribution of each feature is calculated. This process is repeated in various combinations of features. The Shapley value is the average marginal contribution of one feature across all possible combinations of features. It can range from one single feature to all features in the model. Additionally, as we will see further below, SHAP can also produce global explanations.

SHAP is different from LIME in that it uses the original, complex model. In this way, it offers the potential for global interpretability of outputs by aggregating SHAP values across many predictions. Note that this is still a second-hand explanation, as it were. SHAP does not identify the way in which features move through the layers of a neural network. Even less does it identify real causal relations between data points.

Additionally, SHAP allows to identify feature interactions, for instance: How does having two credit cards impact the number of open bills. SHAP calculates this by comparing the effect of all features together against their individual and pairwise effects. Furthermore, SHAP but not LIME, achieves a fair distribution of each importance, because it considers all possible coalitions, calculates the marginal contribution of each feature across these coalitions, and averages these contributions. LIME, by contrast, focuses on local approximations around specific predictions and may fail to capture the true importance of features in the global context of the model.

Applicability in a legal context. What might a use case for these models in a legal context look like? Above, we used the example of a banking supervisor or a consumer advocacy group's interest to receive a detailed explanation of specific feature values. SHAP allows for that. Assume a consumer advocacy group is weighing the odds of a lawsuit based on indirect gender discrimination. One of the test prongs will be to show that the loan applicants are "similarly situated" – you cannot compare apples with pears. The consumer advocacy group might argue as follows: Let's have a look at the subgroup of university professors in the highest income bracket applying for a loan. They would like to know the relative importance of the gender of a university professor in that subgroup. SHAP will give you this value, whereas local models do not allow for contrastive explanations (ibid.: 9.5.4.).

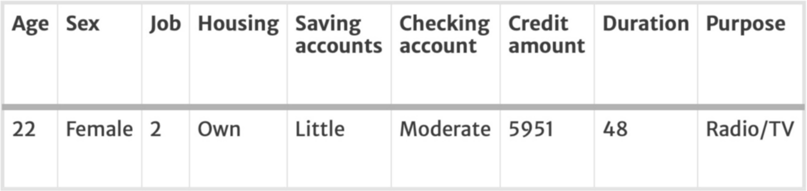
Drawbacks of SHAP. Depending on context, it is important to note that SHAP requires a representative background dataset to avoid unrealistic feature-value combinations. Another disadvantage can be that many versions of SHAP assume feature independence (ibid.: 9.5.5.), which is often unrealistic in an empirical setting. If that assumption breaks down, the model's explanation is less (or not at all) useful. Let us follow up on the consumer advocacy group-example: Many features of individual loan applicants in the subgroup of university professors will be correlated (for the assumption that everything correlates with race, see: Prince and Schwarcz 2020; Langenbucher 2022: 22–27). In fact, the legal doctrine of indirect discrimination was developed to cope with correlated features: If an employer discriminates on the basis of part-time work, he does not directly discriminate against women. However, if the percentage of part-time workers is predominantly female, he discriminates indirectly because gender and part-time work correlate narrowly. Such correlations can influence a model's prediction if two features are highly correlated or if one specific feature is slightly correlated with many features. For SHAP, this can raise important challenges, especially if a truly model-agnostic SHAP is used. Some versions of SHAP cope better than others. TreeSHAP is optimized for tree ensembles like random forest and gradient boosting machines and can better account for feature independence than, for instance, KernelSHAP, even though some extensions to this latter method are being proposed (Shuyang 2024; on yet another approach, Generalized DeepSHAP, see Chen, Lundberg and Lee 2022).

Contrastive and counterfactual explanations. SHAP gives what computer scientists call a "contrastive" explanation. It shows why one specific prediction differs from a baseline, for instance: compared to the average credit score, having seven credit cards lowers the score by 5 points. Additionally, SHAP can compare one prediction to a subset of the data set or even to one single instance, by calculating differences in their feature contributions, for example, scores above or below the threshold a bank sets having to do with income or open bills. This form of transparent, quantitative feature attribution is one of the strengths of SHAP. It answers the question of why a certain outcome (e.g., a score of 10) was reached, instead of a different outcome (e.g., a score of 8). If a user is not so much interested in comparing predictions across instances, but on receiving a recommendation (e.g., get rid of two of your eight credit cards), a counterfactual explanation might be useful. Counterfactual explanations tell us what features must change to produce a certain outcome (e.g., a score) by describing the smallest change to the feature value (e.g., the number of credit cards) that changes the prediction to a predefined output (e.g., a certain threshold score). In that way, counterfactual explanations answer "what if?" questions (Wachter, Mittelstadt and Russell 2018). The goal is to find a set of examples that not only achieve the desired outcome (validity) but are also as close as possible to the original data point (proximity) and differ significantly from each other (diversity) to represent various actionable paths (Mothilal, Sharma and Tan 2020). There are both model-agnostic and model-specific counterfactual explanation methods (Molnar 2022: 15). One explanatory model that delivers counterfactual explanations is DiCE (Mothilal, Sharma and Tan 2020; de Oliveira, Sörensen and Martens 2024). While LIME and SHAP primarily focus on highlighting the importance of individual features for a specific prediction, DiCE proactively generates multiple diverse counterfactual examples to show different ways the outcome could be changed (ibid.; Jain, Sangroya and Vig 2025; Dominici et al. 2025).

Drawbacks of counterfactual models. At first glance, counterfactual explanations, like those generated by DiCE, seem to provide an ideal tool for explaining credit scoring. The potential borrower learns how he can "play" with relevant features, focusing on a small number of changes. However, they generally suffer from several issues concerning their robustness. Minor changes to the underlying model can invalidate the previously generated explanation (Upadhyay, Joshi and Lakkaraju 2021; Hamman et al. 2023). Similarly, slight variations in the input data can lead to entirely different counterfactual suggestions (Slack et al. 2021; Artelt et al. 2021). While methods to enhance robustness exist, they often come with significantly increased computational costs (Jiang et al. 2024). Furthermore, there is no guarantee that the examples generated by DiCE are always realistic or plausible within the data's context or actually feasible for the user to implement (Salimi et al. 2023; Barr et al. 2021), e.g., change your age or gender to receive a better score. Lastly, counterfactual explanations suffer from what computer scientists call the "Rashomon effect", namely that there exist many equally good predictive models for the same dataset (Rudin et al. 2024). Applied to counterfactual models, this translates as receiving multiple different counterfactual explanations, each telling a different story and, possibly, contradicting each other.

Dandl et al. 2020 provide the following example, illustrating their multi-objective counterfactual model (MOC) in the context of credit-scoring:

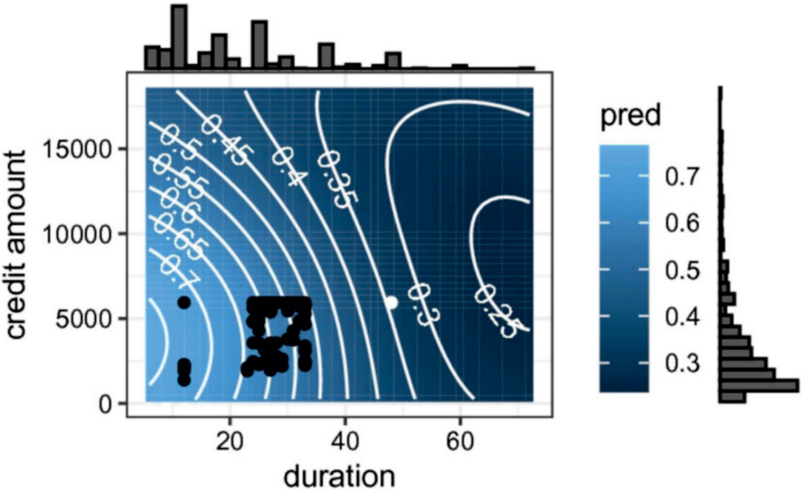
Figure 1: Baseline data of a sample consumer for counterfactual modelling.[1]



Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose
22	Female	2	Own	Little	Moderate	5951	48	Radio/TV

Their model generated a total of 136 counterfactuals and then focused on 82 of them with predictions within [0.5,1]. They produced a response surface plot, suggesting decreasing credit duration and credit amount:

Figure 2: Example for a Response surface plot generated from a counterfactual model.[2]



(b) Response surface plot

Molnar (2022: 15) provides a variation on this example, based on the dataset used in Dandl et al. 2020. The consumer is described as follows:

---

[1] Illustration by Dandl (et al 2020) of their multi-objective counterfactual model (MOC) in the context of credit-scoring. Screenshot from 13.10.2025; used with the authors' consent.

[2] Response surface plot by Dandl (et al 2020), suggesting decreasing credit duration and credit amount. Screenshot from 13.10.2025; used with the authors' consent.

Figure 3: Baseline data for a variation of the first credit scoring example.[3]

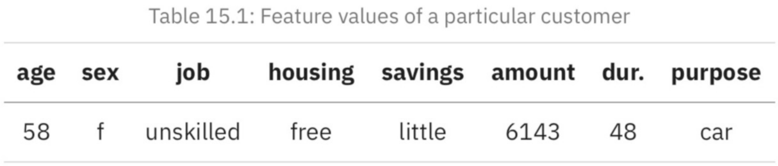


Table 15.1: Feature values of a particular customer

age	sex	job	housing	savings	amount	dur.	purpose
58	f	unskilled	free	little	6143	48	car

The model predicts that the probability that the consumer gets her preferred score is 24.2%. Her interest is to employ a counterfactual explanatory model to understand what she needs to change as to her input features to reach a probability of >50% to get the preferred score. The model displays the following results:

Figure 4: Example counterfactual results illustrating the effects of the suggested changes on the predicted credit score.[4]

Table 15.2: The ten best counterfactuals found for the customer

age	sex	job	amount	dur.	o_2	o_3	o_4	f(x')
		skilled		-20	0.108	2	0.036	0.501
		skilled		-24	0.114	2	0.029	0.525
		skilled		-22	0.111	2	0.033	0.513
-6		skilled		-24	0.126	3	0.018	0.505
-3		skilled		-24	0.120	3	0.024	0.515
-1		skilled		-24	0.116	3	0.027	0.522
-3	m			-24	0.195	3	0.012	0.501
-6	m			-25	0.202	3	0.011	0.501
-30	m	skilled		-24	0.285	4	0.005	0.590
-4	m		-1254	-24	0.204	4	0.002	0.506

---

[3] Variation on the Dandl (et al 2020) example by Molnar (2022, 15), based on the dataset used in Dandl et al (2020). Screenshot from 13.10.2025; used with the author's consent.

[4] Variation on the Dandl (et al 2020) example by Molnar (2022, 15), based on the dataset used in Dandl et al (2020). Screenshot from 13.10.2025; used with the author's consent.

Some of these help: The consumer learns, for instance, that she should lower the duration of the loan. Others are examples of complicated or even unrealistic suggestions: Seven of the ten best counterfactuals suggest to become "skilled", but it is unclear whether the potential borrower has that option. She cannot change her gender to "m" (as suggested by four of the ten best counterfactuals) or lower her age (as suggested by seven of the ten best counterfactuals).

5. Coming Full Circle: Credit Scoring and Explainability

The rough-and-ready overview of different explanatory strategies has highlighted how these work and what some of their advantages and disadvantages are. An important feature to keep in mind is the probabilistic nature of AI systems which accounts for accurate predictions without revealing the underlying causal mechanisms. Often, this produces a disconnect with the law's expectations. Hence, in a legal context, picking the best – or the second best – explanatory strategy depends very much on context.

If a bank asks its financial supervisor to allow it to use a certain model, global explanatory power will matter a lot. By contrast, if a consumer asks for a good-enough, easy-to-understand explanation of his credit score, while the profiler will want to keep his trade secrets, a counterfactual model might be sufficient. For the consumer, it will often be more important to understand his options for behavioral change when confronted with his score. To learn about those, a local explanation that approximates what the complex model does and limits itself to a sparse explanation will often suffice (see Molnar 2022: 9.5.5.).

European Court of Justice in Dun & Bradstreet (D&B). A recent decision by the European Court of Justice nicely illustrates the legal and practical relevance of providing these explanations to AI-based predictions – particularly counterfactual ones that are meaningful to the individual. The case (C-203/22, judgment Feb 27, 2025) involved a plaintiff who was denied a mobile phone contract based on an opaque credit score provided by D&B, who subsequently refused to disclose a detailed explanation of the underlying computation, citing trade secrets (Langenbucher and Bauer 2025). In interpreting the requirement of "meaningful information about the logic involved" under Art. 15(1)(h) GDPR, the ECJ clarified that a credit scoring company is not required to provide complex algorithms. Instead, the Court emphasized the need for "clear, understandable explanations", seen from an average consumer's point of view. Crucially, the Court suggested that explaining how changes to the individual's data would have led to a different score could satisfy this requirement – an approach strongly aligning with the concept of counterfactual explanations (ibid.). Methods like DiCE, designed to generate diverse, actionable counterfactuals (e.g., "If the consumer had one less credit card…"), thus present a potential technical solution for fulfilling these transparency obligations while respecting intellectual property rights (ibid.).

However, this raises further practical questions: While the ECJ endorsed the possibility of courts reviewing information, concerns remain about whether courts possess the necessary technical expertise to adequately assess the validity and potential manipulation of counterfactuals generated solely by the scoring entity. This challenge suggests a likely need for independent technical experts or neutral intermediaries to verify the reliability and completeness of such explanations in practice (ibid.).

XAI methods – from local approximations like LIME, game-theoretic approaches like SHAP, to counterfactual explanations using DiCE – offer various tools to make the functioning of AI systems more understandable. However, as the analysis of legal requirements from GDPR, the AI Act, and more specific regulations like the German *Bundesdatenschutzgesetz*-draft shows, explainability often serves as a vehicle to realize core aspects of the "trustworthiness" sought by the legislator – such as traceability, fairness, accountability, and the possibility of effective redress. The choice of the "right" explanation method depends heavily on context-specific needs: While a bank supervisor might wish to gain access to global insights (SHAP), an affected consumer might be looking for concrete options for action (DiCE, MOC). The challenge for the future lies in integrating these technical possibilities with the legal framework, combining usability in practice with risk-awareness for all parties concerned.

References

Aljohani, Manar et al. (2025): "A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare", arXiv:2502.15871.

Artelt, André et al. (2021): "Evaluating Robustness of Counterfactual Explanations", in: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8, arXiv.2103.02354.

Barocas, Solon and Selbst, Andrew (2016): "Big Data's Disparate Impact", in: California Law Review 104(3), pp. 671–732.

Barr, Kyle et al. (2021): "Counterfactual Explanations via Latent Space Projection and Interpolation", arXiv:2112.00890.

Breuer, Janice and McDermott, John (2008): "Trustworthiness and Economic Performance", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1314844.

Chen, Hugh, Lundberg, Scott M. and Lee, Su-In (2022): "Explaining a Series of Models by Propagating Shapley Values", in: Nature Communications 13, 4512.

Dandl, Susanne et al. (2020): "Multi-Objective Counterfactual Explanations", https://link.springer.com/chapter/10.1007/978-3-030-58112-1_31.

Dominici, Gabriele et al. (2025): "Counterfactual Concept Bottleneck Models", https://openreview.net/forum?id=w7pMjyjsKN.

Dubovitskaya, Elena and Bosold, Gregor (2025): "Right of Explanation of AI Decisions" [Beitrag in diesem Band].

Freiesleben, Timo (2022): "What Does Explainable AI Explain", Dissertation, LMU Munich, https://edoc.ub.uni-muenchen.de/31933/1/Freiesleben_Timo.pdf.

Hamman, Matthew et al. (2023): "Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees", in: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, pp. 12384–12401.

Jain, Suparshva, Sangroya, Amit and Vig, Lovekesh (2025): "DifCluE: Generating Counterfactual Explanations with Diffusion Autoencoders and Modal Clustering", in: Proceedings of ACM Conference, New York, USA, arXiv:2502.11509v1.

Jiang, Ziyi, Leofante, Francesco, Rago, Antonio and Toni, Francesca (2024): "Robust Counterfactual Explanations in Machine Learning. A Survey", arXiv:2402.01928.

Kaminski, Andreas (2025): "Trust in AI. A Unified Approach" [Beitrag in diesem Band].

Langenbucher, Katja (2022): "Consumer Credit in The Age of AI. Beyond Anti-Discrimination Law", ECGI Law Working Paper 663, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4298261.

Langenbucher, Katja (2024): "Financial Profiling", Conference on Mapping and Governing the Online World; Ascona, Switzerland, https://www.researchgate.net/publication/381127142_Financial_Profiling.

Langenbucher, Katja and Bauer, Kevin (2025): "Explaining Credit Scores. The ECJ Rules on Automated Credit Assesments", Compliance & Enforcement, PCCE at NYU, https://wp.nyu.edu/compliance_enforcement/2025/03/18/explaining-credit-scores-the-ecj-rules-on-automated-credit-assessments/, last access: June 18, 2025.

Molnar, Christoph (2022): Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/.

Mothilal, Ramaravind K., Sharma, Amit and Tan, Chenhao (2020): "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations", in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Association for Computing Machinery, New York, NY, USA, pp. 607–617.

Nisevic, Maja, Cuypers, Arno and de Bruyne, Jan (2025): "Explainable AI. Can the AI Act and the GDPR go out for a Date?", International Joint Conference on Neural Networks, Yokohama, Japan, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5056022.

de Oliveira, Raphael M. B. de, Sörensen, Kenneth and Martens, David (2024): "A Model-agnostic and Data-independent Tabu Search Algorithm to Generate Counterfactuals for Tabular, Image, and Text Data", European Journal of Operational Research 317(2), pp. 286–302.

Prince, Anya and Schwarcz, Daniel (2020): "Proxy Discrimination in the Age of Artificial Intelligence and Big Data", Iowa Law Review 105, pp. 1257–1318.

Ribeiro, Marco Tulio, Singh, Sameer and Guestrin, Carlos (2016): "Why Should I Trust You? Explaining the Predictions of any Classifier", in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1135–1144, California, USA, arXiv:1602.04938.

Rudin, Cynthia et al. (2024): "Amazing Things Come From Having Many Good Models", arXiv:2407.04846v1.

Shuyang, Xiang (2024): "KernelSHAP Can Be Misleading With Correlated Predictors", TDS Archive, https://towardsdatascience.com/kernelshap-can-be-misleading-with-correlated-predictors-9f64108f7cfb/.

Salimi, Pedram et al. (2023): "Towards Feasible Counterfactual Explanations. A Taxonomy Guided Template-based NLG Method, Frontiers in Artificial Intelligence and Applications", 372, pp. 2057–2064.

Slack, Dylan et al. (2021): "Counterfactual Explanations Can Be Manipulated", in: Advances in 34 Neural Information Processing Systems (NeurIPS 2021), https://proceedings.neurips.cc/paper/2021/hash/009c434cab57de48a31f6b669e7ba266-Abstract.html.

Solove, Daniel J. (2024): "Data Is What Data Does. Regulating Based on Harm and Risk Instead of Sensitive Data", in: Northwestern University Law Review 118, pp. 1081–1138.

Upadhyay, Samanvay, Joshi, Shweta and Lakkaraju, Himabindu (2021): "Towards Robust and Reliable Algorithmic Recourse", in: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp. 29536–29548.

Wachter, Sandra, Mittelstadt, Brent and Russell, Chris (2018): "Counterfactual Explanations Without Opening the Black Box. Automated Decisions and the GDPR", in: Harvard Journal of Law & Technology 31, pp. 841–887.

Zhang, Zhen et al. (2024): "Visual Analysis of Trustworthiness Studies. Based in the Web of Science Database", in: Front. Psychol 15, 1351425.

Legal Sources

Court of Justice of the European Union, First Chamber. Case C-203/22 Magistrat der Stadt Wien v Dun & Bradstreet Austria GmbH. Judgment of 27 February 2025, ECLI:EU:C:2025:117.

European Commission (2018): "Communication from the Commission, Artificial Intelligence for Europe", COM(2018) 237 final, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237.

European Commission (2019): "Ethics Guidelines for Trustworthy AI, report", https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Commission (2020): "On Artificial Intelligence. A European Approach to Excellence and Trust, White Paper", COM(2020) 65 final, https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, (GDPR).

Directive (EU) 2023/2225 of the European Parliament and of the Council of 18 October 2023 on Credit Agreements for Consumers, (CCD).

Deutscher Bundestag, Drucksache 20/10859 of 27 March 2024, Draft of the First Act to Amend the Federal Data Protection Act (BT- Drucksache 20/10859).

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, (AI Act).