

We Need Explanation Cards to Connect Explanation Algorithms to the Real World

Eric Günther^{1,2}, Balázs Szabados³, Kristof Meding¹, Gunnar König^{1,2}, Sebastian Bordt^{1,2}, Ulrike von Luxburg^{1,2},

¹University of Tübingen

²Tübingen AI Center

³HUN-REN Institute for Computer Science and Control (SZTAKI), Budapest, Hungary

{eric.guenther, kristof.meding, gunnar.koenig, sebastian.bordt, ulrike.luxburg}@uni-tuebingen.de, balazs.szabados@sztaki.hu

Abstract

Algorithmic explanations are intended to help stakeholders understand opaque algorithmic decisions, but in practice, they often fall short. First, the meaning of algorithmic explanations is often not what one might intuitively expect, so expert knowledge is required to interpret them correctly. Second, recent work has shown that popular explanation algorithms are uninformative about the behavior of complex decision functions. Together, these issues create a gap between what explanations appear to convey and what they actually provide. In this work, we propose *Explanation Cards for Explanation Algorithms*, which augment standard explanations with complementary information about robustness and validity, as well as clear instructions for interpretation. The complementary information can render otherwise uninformative explanations practically useful, while also helping to detect cases where they are not. Importantly, the interpretation instructions in explanation cards shift responsibility from users to providers: Rather than expecting users to recognize what can and cannot be concluded from an explanation, providers must make this explicit upfront. Using counterfactual explanations and SHAP as examples, we demonstrate how providers can construct explanation cards and that these cards provide users with the guidance needed for sound interpretation. We further argue that explanation cards might offer a practical way to operationalize some provisions of the EU AI Act. Overall, explanation cards are a significant step toward making explanation algorithms fit for real-world use cases.

Code — <https://github.com/tml-tuebingen/explanation-card>

Version with appendix — <https://arxiv.org/abs/2606.16786>

1 Introduction

Explainable machine learning (XAI) methods are meant to help stakeholders understand opaque, black-box machine learning models. For example, XAI methods have been discussed as a tool towards enabling transparency for users affected by algorithmic decisions, and they have also been discussed as tools to debug algorithmic decisions (Ribeiro, Singh, and Guestrin 2016; Wachter, Mittelstadt, and Russell 2017). Meanwhile, the field of explainable machine learning faces fundamental criticism, and various authors have advocated against their use in high-stakes settings (Lipton 2018;

Rudin 2019; Krishnan 2020; Krishna et al. 2022; Freiesleben and König 2023; Corbett and Denton 2023; Schmitt et al. 2024; Deck et al. 2024; Skirzynski, Danks, and Ustun 2025; Bordt, Raidl, and von Luxburg 2025).

Challenge of Misinterpretation. A prominent criticism is that algorithmic explanations lend themselves to misinterpretation (Ghassemi, Oakden-Rayner, and Beam 2021; Bordt, Raidl, and von Luxburg 2025), leading stakeholders to derive wrong conclusions from explanations (Molnar et al. 2020; Bordt et al. 2022; Freiesleben and König 2023; Huang and Marques-Silva 2023; Geirhos et al. 2024; Tomaszewska and Biecek 2024; Khadar et al. 2025). One cause of misinterpretation is that the meaning of explanations is often not what one would intuitively expect (Bhatt et al. 2020). Instead, detailed expert knowledge is required to translate them into justified insight (Khadar et al. 2025). This is particularly problematic in high-stakes settings, where explanations are intended to help non-expert stakeholders understand how a machine learning model arrived at a particular decision (Bordt et al. 2022). For example, Ghassemi, Oakden-Rayner, and Beam (2021) notes in a medical context that the “interpretability gap of explainability methods relies on humans to decide what a given explanation might mean”.

Technical Challenges of Explaining Complex Models.

Research on the technical properties of explanation algorithms has highlighted significant drawbacks in many popular algorithms. For example, it has been shown that the output of explanation algorithms is sensitive to seemingly irrelevant changes in the model that one aims to explain (Dombrowski et al. 2019; Slack et al. 2020). More generally, work has shown that strong assumptions about the underlying model are required for explanations to be practically useful, and that the explanations themselves provide no means of assessing whether those assumptions are met (Garreau and von Luxburg 2020; Bordt, König, and von Luxburg 2023). In particular, Günther et al. (2025) has demonstrated that local feature attributions are only informative about models that are sufficiently simple. Taken together, these results suggest that in settings where the model is complex, algorithmic explanations may not justify any intuitive interpretation at all.

Explanation Cards for Explanation Algorithms. In this paper, we propose a pragmatic solution that addresses both the problem of model complexity and misinterpretation:

Explanation Cards for Explanation Algorithms. Explanation Cards augment the output of standard explanation algorithms with complementary information about their robustness and validity, and provide clear instructions for how the output of the explanation algorithms should be interpreted. From a technical perspective, this complementary information renders the explanations informative. From the perspective of the stakeholder who receives the explanation card, the complementary information makes the explanation practically useful whenever possible, while also helping to detect cases where it is not. Most importantly, the instructions for interpretation shift responsibility from the user to the provider of the explanation: Rather than expecting the user to recognize what can and cannot be concluded from an explanation, providers must make this explicit upfront.

Our key contributions are the following:

- We suggest using explanation cards to alleviate challenges in interpreting the output of common explanation algorithms. To construct such cards, one needs to derive technical guarantees about the specific explanation, and translate those to natural-language statements that outline to a user which concrete interpretations can and cannot be drawn from the given explanation.
- We derive the necessary mathematical guarantees to construct two example explanation cards: an explanation card for counterfactual explanations for loan decisions (Section 4), and an explanation card for SHAP explanations in medical diagnosis (Section 5).
- In a legal analysis, we show how the use of explanation cards could help to satisfy the transparency and explainability standards set by the EU AI Act (Section 6).

2 Related Work

In this paper, we focus on post-hoc explanation algorithms (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Wachter, Mittelstadt, and Russell 2017). They are widely used, for example, for model debugging (Adebayo et al. 2020), but also in high-stakes scenarios like medicine (Borys et al. 2023), recourse recommendations (Karimi et al. 2022) or detecting discrimination (Agarwal et al. 2018). For a general overview of the big field of explainable AI (XAI), we refer the reader to Molnar (2025).

Counterfactual Explanations. The first explanation algorithm discussed in this work is counterfactual explanations (Wachter, Mittelstadt, and Russell 2017). Counterfactual explanations have gained significant attention because they align with how humans naturally reason about cause and effect (Miller 2019; Byrne 2019). They have also been proposed as a recourse mechanism, for example, in credit lending (Ustun, Spangher, and Liu 2019). However, it is also known that counterfactuals lack robustness: An adversary can train models with unstable counterfactuals (Slack et al. 2021) and generally, small changes in the observation can lead to different explanations (Fokkema, De Heide, and Van Erven 2023). Furthermore, changes in one feature may have unexpected downstream causal effects on

other features (Karimi, Schölkopf, and Valera 2021; Karimi et al. 2022; König, Freiesleben, and Grosse-Wentrup 2023; König et al. 2025) and a counterfactual may be invalidated if changed slightly (Dominguez-Olmedo, Karimi, and Schölkopf 2022). More recently, it was shown that counterfactuals are not informative for arbitrarily complex functions (Günther et al. 2025). These insights have led to a number of attempts to make counterfactuals more robust (Artelt et al. 2021; Zhang et al. 2023; Leofante and Potyka 2024). Particularly related to our work are regional counterfactuals (Fernández et al. 2020; VanNostrand et al. 2023), which have been proven to be easier to understand for a subject (VanNostrand et al. 2024).

SHAP Explanations. The second explanation algorithm discussed in this work is SHAP (Lundberg and Lee 2017). SHAP has become one of the most widely used post-hoc explanation algorithms due to its ease of use and model-agnostic applicability (Lundberg, Erion, and Lee 2018; Molnar 2025). However, the algorithm has also been frequently criticized (Kumar et al. 2020): Similar to counterfactual explanations, SHAP is sensitive to off-manifold behavior (Frye et al. 2020; Taufiq, Blöbaum, and Minorics 2023) and can be manipulated by an adversary (Slack et al. 2020; Hwang et al. 2025). Furthermore, the presence of variable interactions limits the interpretation of the SHAP attributions, which has given rise to algorithmic variants that also attribute interactions (Lundberg, Erion, and Lee 2018; Zhang et al. 2021; Bordt and von Luxburg 2023; Fumagalli et al. 2023). Techniques like these also exist for global methods that aim to explain the data (König, Günther, and von Luxburg 2025). But even for models without interactions, SHAP values remain uninformative for complex functions due to their potential instability (Günther et al. 2025). We hence present several SHAP explanations, presenting their stability, and give additional information that reveals whether they connect to the prediction, which is only possible if the model does not exhibit strong interactions.

Legal Discussion Around Explainability. In the context of high-stakes decisions, the connection to the legal domain has also been widely examined: Explainability requirements have been discussed in relation to the General Data Protection Regulation (GDPR) (Wachter, Mittelstadt, and Russell 2017; Hamon et al. 2022; Selbst and Powles 2018), various drafts and the final version of the AI Act (Panigutti et al. 2023; Kaminski and Malgieri 2025), and their combined implications (Hacker and Passoth 2020; Ebers 2020; Nannini 2024).

Explanation Cards are inspired by model cards (Mitchell et al. 2019) and datasheets (Gebru et al. 2021). They are also closely related to saliency cards, proposed by Boggust et al. (2023). However, we develop explanation cards for tabular data rather than for images, and tailor them to a non-expert audience. Explanation cards move beyond previous approaches that enhance explanation algorithms with robustness and stability measures (Bhatt, Weller, and Moura 2021; Velmurugan et al. 2021; Dasgupta, Frost, and Moshkovitz 2022; Dhurandhar et al. 2024) because they both enrich the explanations with additional information *and* focus on how this information can be communicated to

stakeholders in real-world applications. In this respect, explanation cards provide a step towards user-centric explainability (Miller 2019; Liao and Varshney 2021; Freiesleben and König 2023; Schor et al. 2024; Hullman, Guo, and Ustun 2025).

3 Problem Setup

We consider machine learning algorithms that make predictions in social contexts. Specifically, we consider the setup where an algorithm is trained on **tabular data** to predict individual characteristics, such as a medical diagnosis or the probability of loan repayment. Each dimension of the input encodes a different property of a person, for example, their age, income, or blood pressure. The machine learning algorithm is used to learn a **decision function** $f : \mathbb{R}^d \rightarrow \mathbb{R}$, outputting a score or a probability $f(x_0)$ for an individual x_0 . This score can be transformed to a binary prediction (“receives the loan” or “does not receive the loan”). In addition to the prediction, the provider of the machine learning system provides an explanation. In this work, we consider **counterfactual explanations**, introduced by Wachter, Mittelstadt, and Russell (2017), and **SHAP**, introduced by Lundberg and Lee (2017). Counterfactual explanations provide another datapoint close to x_0 that has a different prediction. SHAP provides a scalar feature importance value for each dimension of the input.

Our main research question is whether individuals who are affected by machine predictions can make use of the explanations to answer questions that are **practically relevant** (Liao and Varshney 2021; Bordt, Raidl, and von Luxburg 2025). For example, a loan applicant might want to know whether earning more would be sufficient to be granted the loan, and a doctor might want to know how a change in a patient’s blood pressure would change the predicted outcome.

We also ask whether the provided explanations satisfy the **informativeness** condition from Günther et al. (2025). Informativeness is a mathematical notion that quantifies whether an explanation provides any information about the behavior of the decision function. Informativeness can be seen as a necessary condition on explanation algorithms, as a non-informative explanation does not provide a robust description of the predictive behavior of the decision function. A key insight of this framework is that explanations tend to be informative if one can construct local regions in which the explanations behave benignly, for example, they only vary by a small amount. These regions will show up on our explanation cards below, for example, as “region of stability” in the case of counterfactuals.

In the mathematical parts of the paper, we consider a dataspace $\mathcal{X} \subseteq \mathbb{R}^d$ with a probability distribution P . For the set of all features, we write $[d] := \{1, \dots, d\}$. We denote the features of a point x by $(x^{(1)}, \dots, x^{(d)})$. If we consider a subset of features $S = \{j_1, \dots, j_{|S|}\} \subseteq [d]$ of some point $x \in \mathcal{X}$, we write $x^{(S)} := (x^{(j_1)}, \dots, x^{(j_{|S|})})$. We aim to explain the prediction of the function f at the point $x_0 \in \mathcal{X}$. We assume that f is a real-valued scoring function from which we can derive a final binary prediction by thresholding, t denotes the threshold with respect to which the prediction is derived.

4 Explanation Cards for Counterfactual Explanations

In this section, we introduce Explanation Cards for counterfactual explanations. We begin by discussing the intuitive interpretations that humans may assign to counterfactual explanations (Section 4.1). We then show that these interpretations are invalid for decision functions that have a complex decision boundary (Section 4.2). Explanation Cards can overcome this problem: They provide the user with an interpretation that is valid for the given decision function (Section 4.3).

4.1 Interpreting Counterfactual Explanations

We begin by discussing the real-world interpretations that humans may assign to counterfactual explanations. For illustration, we consider a simplified setup where a machine learning model determines loan approval based on three attributes: age, income, and current account balance (Wachter, Mittelstadt, and Russell 2017; Mothilal, Sharma, and Tan 2020; Karimi, Schölkopf, and Valera 2021; Verma et al. 2024). As an affected person, the loan applicant has under certain circumstances the right to obtain clear and meaningful explanations according to Article 86 of the European AI Act. Concretely, assume that a 32-year-old applicant with a yearly income of \$50,000 and a current account balance of \$4,000 is denied a loan. To offer recourse, the bank provides a counterfactual explanation, a nearby datapoint from a different class.

An automated system has denied your loan application because your current account balance is too low. To receive the loan upon reapplication, you are suggested to increase your account balance to \$14,000.

The applicant now attempts to implement the counterfactual explanation. To do so, she will *interpret* the counterfactual explanation, meaning that she attempts to infer real-world behaviors that would lead to loan approval (Wachter, Mittelstadt, and Russell 2017; Ustun, Spangher, and Liu 2019; Karimi, Schölkopf, and Valera 2021; Bordt, Raidl, and von Luxburg 2025). For example, the applicant may conclude that increasing her yearly income to \$55,000 will allow her to save \$5,000 per year, so that at age 34, she will have saved the required \$10,000. In doing so, the applicant makes implicit assumptions about the explanation and the machine learning algorithm’s behavior. Concretely, the applicant assumes that:

1. **Monotonicity:** Although the explanation requires an increase of \$10,000 in balance, a growth of more than \$10,000 would also result in a positive outcome. In general, the counterfactual would still be valid if the yearly income and account balance are increased further.
2. **Stability of the Counterfactual:** The provided explanation will still be valid if, over the course of implementing the explanation in the real world, the values of some of the other features change. In our example, the applicant assumes that the counterfactual explanation remains valid even after she has become two years older.

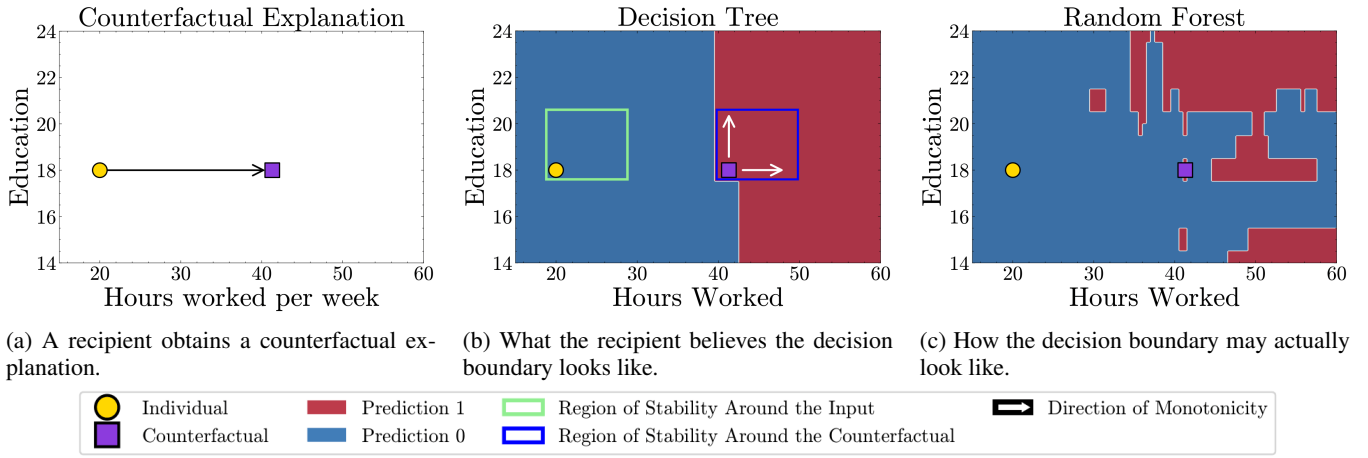


Figure 1: **Intuitive Interpretations of Counterfactual Explanations May Be Invalid for Complex Models.** The left panel illustrates a **data point** and its **counterfactual explanation**. It summarizes the factual knowledge that a user has after receiving their explanation. The middle plot illustrates a decision boundary as it might implicitly be assumed by the user who tries to interpret the received explanation. It satisfies the properties of monotonicity (white arrows), stability around the counterfactual (blue box), and stability around the input (green box). The right panel shows the decision boundary of a complex model as it might be used by the bank. Here, the counterfactual does not satisfy stability or monotonicity. For this model, the user’s intuitive interpretations are invalid.

In addition, we argue that the applicant might also assume that

3. **Stability Around the Input:** Other applicants with a similar background are suggested to implement similar changes. For example, people in the 30-35 age group, with a salary of \$45,000-55,000, and a balance lower than \$5,000, would be granted the loan after increasing their savings by approximately \$10,000.

The fact that the applicant draws conclusions about the real world from the explanations is not specific to counterfactuals: Work in human-centered explainability has shown that such interpretations are what human decision makers are ultimately interested in (Liao and Varshney 2021; Raes et al. 2024).

4.2 Intuitive Interpretations of Counterfactual Explanations May Be Invalid for Complex Models

Whether the intuitive interpretations from the previous Section 4.1 are valid depends strongly on the geometry of the decision boundary – that is, on the specific machine learning model that the bank employs (Laugel et al. 2019; Artelt et al. 2021; Zhang et al. 2023). This is illustrated in Figure 1. In the left panel of the Figure 1, we illustrate an individual and a corresponding counterfactual explanation. In the middle panel of Figure 1, we illustrate the decision boundary as it might be implicitly imagined by a user who interprets the explanation as discussed in the previous Section 4.1. For this decision boundary, the user’s intuitive interpretations would be valid. In the right panel of Figure 1, we depict how the decision boundary of the machine learning model may actually look like. Concretely, we depict the decision boundary

of a random forest classifier, a popular machine learning algorithm for tabular datasets. For this decision boundary, the user’s intuitive interpretations are invalid: The counterfactual is not monotone, because moving further into the hours worked direction invalidates the counterfactual. The counterfactual is also not stable around the input, because slightly increasing the hours worked moves the counterfactual in the direction of the negative class. Similarly, the counterfactual itself is not stable.

It is important to stress that the user who receives the explanation cannot know whether the bank’s function behaves as in the middle plot (which would be good) or as in the right plot (in which case the explanation, albeit formally being a correct counterfactual, might be misleading to the user). Of course, one could argue that a bank might be unlikely to use a funny decision function as in the right plot. But the key point is that the user cannot take this for granted, which directly leads to our explanation cards: Whether the bank uses a nice model, as in the middle of the plot, or a funny model, as in the right plot, will become apparent from the explanation card that we introduce below.

Whether a counterfactual explanation is stable also has strong mathematical implications. We can formalize the above conditions as mathematical assumptions on the explanations. In each of them, t is the decision threshold, meaning all points $x \in \mathcal{X}$ yielding $f(x) < t$ are assigned the same label as x_0 , and all points $x \in \mathcal{X}$ yielding $f(x) \geq t$ are assigned the other label.

1. **Monotonicity:** Let x_C be a counterfactual explanation, and let $\mathbb{X} \subseteq \mathcal{X}$ be such that $x_C \in \mathbb{X}$ and $\forall x \in \mathbb{X} : f(x) \geq t$. The set $\mathbb{X}$ is called monotonically increasing in feature j if for all $x \in \mathbb{X}$ it holds that $f(x + \lambda e_j) \geq t$ for any $\lambda \geq 0$, where e_j is the j -th standard basis vector.

An automated system has denied your loan application because you do not have savings. To receive a loan upon reapplication, you are suggested to open a savings account and increase your savings to 100 DM.

The action suggested by the counterfactual explanation remains valid as long as you do not open a checking account, and your savings are *above* 100 DM. Your housing and possible other credits do not influence the prediction.

However, changes in other features, such as your job, the number of people financially dependent on you, etc. may alter the loan decision. In these cases, the reapplication may no longer be successful.

Similar applicants who are younger than 26.5 years, do not have a checking account, and have savings of less than 100 DM are denied the loan for the same reason and are advised of the same changes.

Figure 2: Counterfactual. Stability of the Counterfactual. Stability around the input. Further caveat.

Explanation Card for a Counterfactual Explanation in Loan Approval. The explanation card provides the machine decision, the counterfactual, as well as the necessary information about the stability that is required for interpretation. It also provides information about the *monotonicity* of the counterfactual. In this example, the classifier is a decision tree trained on the South German Credit dataset (Grömping 2019).

2. **Stability of the Counterfactual:** A counterfactual explanation x_C is stable if there exist vectors $l, u \in \mathbb{R}^d$ such that $l < x_C < u$ and $f(x) \geq t$ for all $l < x < u$, where the inequalities are coordinate-wise and $\mathbb{R} := \mathbb{R} \cup \{-\infty, \infty\}$.
3. **Stability Around the Input:** A counterfactual explanation x_C for the data point x_0 is stable around the input if there exist vectors $l, u \in \mathbb{R}^d$ such that $l < x_0 < u$ and $f(x) < t \leq f(x + (x_C - x_0))$ for all $l < x < u$. Again, the inequalities are coordinate-wise and $\mathbb{R} := \mathbb{R} \cup \{-\infty, \infty\}$.

Now, we can show that counterfactuals that do not provide a region of stability do not provide reliable information about the behavior of the decision function:

Theorem 1 (Non-Intuitive Counterfactual Explanations Are Not Informative, Informal). *Let $x_0 \in \mathcal{X}$ be such that $f(x_0) < t$, and let $x_C \in \mathcal{X}$ be a counterfactual explanation for x_0 . If x_C is neither a stable counterfactual nor stable around the input, then it is non-informative. The monotonicity of x_C alone, that is, $\mathbb{X} = \{x_C\}$ is not sufficient for informativeness.*

For a discussion of the mathematical concept of “informative” explanations, we refer to Günther et al. (2025); a formal statement of the theorem and the proof can be found in Appendix A.1 of the arXiv version.

4.3 Explanation Cards Provide the User with Valid Interpretations

In this section, we propose an explanation card for counterfactual explanations. The explanation card provides information about the counterfactual’s stability, enabling the user to interpret it. In a nutshell, **the provider of the explanation, who has access to the decision boundary, enriches the explanation with additional information and**

clearly communicates to the user which interpretations are valid.

Figure 2 depicts our suggestion for an explanation card for a counterfactual explanation. In the first paragraph, the explanation card provides the machine decision and the counterfactual explanation. In the second paragraph, the explanation card provides information about the stability of the counterfactual: In the provided example, it communicates to the user that the counterfactual remains valid as long as they sufficiently increase their savings and do not open a checking account. The third paragraph provides the applicant with additional information about how changes in other features, for example stemming from downstream causal effects of the implemented action, may affect the credit decision. In the last paragraph, the explanation card provides additional information about the stability around the input: Individuals in the same age group who have a similar financial background obtain the same prediction and suggested change as the applicant. Additional examples of explanation cards for counterfactuals are provided in Appendix B of the arXiv version.

At an intuitive level, the benefits of the additional information in the explanation card are obvious: The applicant now knows what conclusions can and cannot be drawn from the explanations. As we show in the following theorem, these benefits can also be described mathematically:

Theorem 2 (Stable Counterfactual Explanations Are Informative, Informal). *Let $x_0 \in \mathcal{X}$ be such that $f(x_0) < t$, and let $x_C \in \mathcal{X}$ be a counterfactual explanation for x_0 . If x_C is a stable counterfactual or stable around the input, then it is informative.*

The formal theorem and its proof can be found in the Appendix A.2 of the arXiv version. An important property of the explanation card is that it provides the size of the regions within which the counterfactual remains stable. Mathematically, it is possible that the counterfactuals are stable only to

tiny changes in the input variables; this would probably be considered useless by most people. Explanation cards reveal this and make the actual amount of stability transparent to the user.

5 Explanation Cards for SHAP

As follows, we propose Explanation Cards for SHAP, one of the most popular explanation algorithms (Lundberg and Lee 2017). SHAP decomposes the prediction at point x_0 into a sum of individual contributions of each feature, plus the expected prediction as a baseline. Mathematically, this means

$$f(x_0) = \Phi_1(x_0) + \dots + \Phi_d(x_0) + E(f(X)),$$

where $\Phi_j(x_0)$ is the SHAP value of feature j at x_0 . For a detailed mathematical definition, see Appendix C.1 of the arXiv version. We consider the common interventional version of SHAP, which is also implemented in the Python package `shap` (Lundberg and Lee 2017; Janzing, Minorics, and Blöbaum 2020).

In Section 5.1, we consider a hypothetical example where a doctor uses SHAP to understand a model and derive intuitive interpretations they might draw from the SHAP values. As we will see in Section 5.2, these interpretations rely on assumptions about the underlying model, which may not hold in complex decision models. When presented with the corresponding explanation card, introduced in Section 5.3, the doctor is able to assess whether their intuitive interpretations are justified.

5.1 Intuitive Interpretations of SHAP Values

Consider the following scenario: A doctor is using an AI model to predict whether a patient has chronic heart disease. As such, the doctor could be a deployer in the sense of Article 3(4) of the AI Act and the provider of the system might need to ensure transparency according to Article 13 AIA. Assume the doctor is given SHAP explanations to gain insight into the model’s prediction behavior.¹ For example, suppose the patient in question is 43 years old, and the corresponding SHAP value for age is -0.22 . From this single SHAP value, there may arise the first misinterpretation:

1. **Stability:** Similar patients of a similar age have a similar SHAP value.

This interpretation seems natural. However, it does not necessarily hold. A common way to get a more representative picture, is by plotting several SHAP values in a SHAP dependence plot. It shows SHAP values across many instances by age, allowing for a better understanding of their stability. Figure 3 shows such a dependence plot across the dataset, confirming that most patients around age 43 have similar negative SHAP values, ranging from roughly -0.3 to -0.1 .

¹Note that in this setting, the doctor does *not* aim to learn about real-world relationships between the features and the target. Rather, the doctor wants to decide whether to trust the model’s decision. Assuming that interventional SHAP provides insight into real-world statistical or causal relationships is another common misinterpretation; see Chen et al. (2020); Janzing, Minorics, and Blöbaum (2020); Freiesleben et al. (2024) for a discussion.

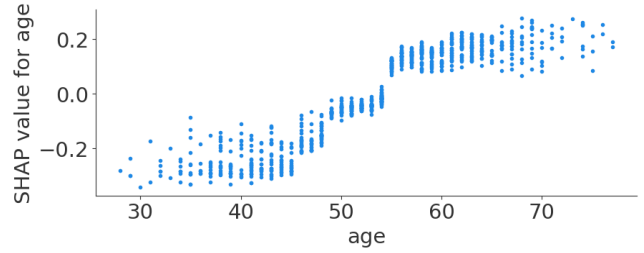


Figure 3: A classical SHAP dependence plot for age, generated by the `shap` package.

However, even this SHAP dependence plot might still lead to misinterpretation. In particular, the doctor may conclude:

2. **Extrapolation:** The SHAP value of the 43-year-old patient is negative and the one of a 53-year-old patient is zero. Hence, the patient would have a higher predicted risk if they were older.
3. **Monotonicity:** The SHAP values in the plot increase with age. Hence, the predicted risk increases with age.

In general, neither of these interpretations holds. In the following section, we explain why and subsequently address the problem using our explanation card in Section 5.3.

5.2 The Intuitive Interpretations May Be Invalid for Complex Models

The presented interpretations do not hold for arbitrarily complex models: Interpretation 1 needs the model to be stable, and Interpretation 2 and 3 need the model to be additive. In the following, we formalize the interpretations, discuss them further, and connect them to mathematical theory.

SHAP Values Are Stable Only if the Model is Stable
We begin by formalizing the concept of stability for SHAP values.

Definition 3 (Local Stability). *For an instance $x_0 \in \mathbb{R}^d$, we call its SHAP values **locally stable** with respect to a local region U around x_0 and tolerance thresholds $\varepsilon_1, \dots, \varepsilon_d$ if $|\Phi_j(x) - \Phi_j(x_0)| \leq \varepsilon_j$ for $j = 1, \dots, d$ and all $x \in U$.*

It is clear that the SHAP values can only be stable if the model itself is stable: As it holds $f(x) = \Phi_1(x) + \dots + \Phi_d(x) + E(f(X))$, the model can vary by at most $\varepsilon_1 + \dots + \varepsilon_d$ within a region U if our SHAP values are stable within U with respect to tolerance thresholds $\varepsilon_1, \dots, \varepsilon_d$. As one is usually not given any stability guarantees for a model, the stability of the SHAP values is not guaranteed either. As a counterexample, consider Figure 4a. It shows the SHAP values for age of a patient (red dot) as well as their 50 nearest neighbors (blue dots, representing a local region U). As we see, there is no stability in the SHAP value, not even for those neighbors of the same or very similar age, who have SHAP values with positive as well as negative signs. Hence, it is always important to consider several SHAP values, for example, in a SHAP dependence plot, if one wants to understand the prediction mechanism of the model. A single SHAP value is not representative of that.

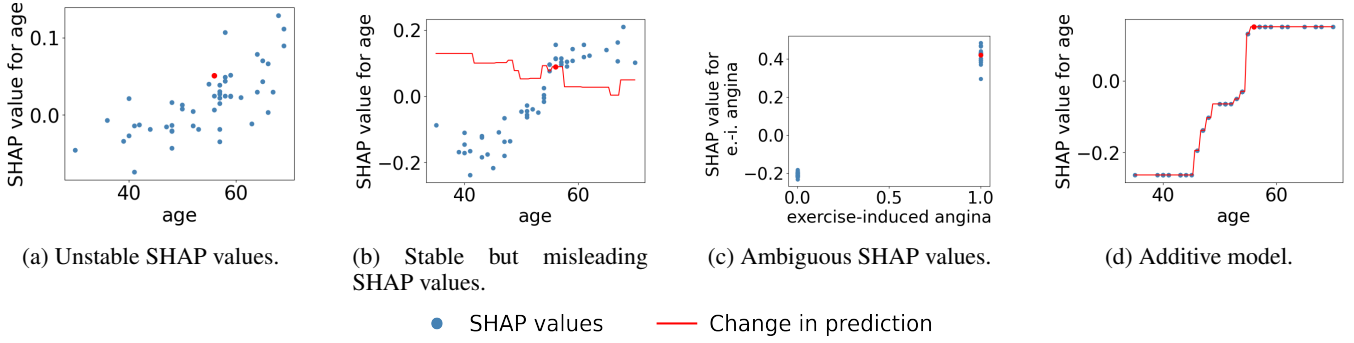


Figure 4: SHAP dependence plots for one datapoint (red) and its 50 nearest neighbors in the input space (blue). Figures 4b and 4d additionally show the prediction change when varying age while keeping all other features fixed. Figure 4a illustrates unstable SHAP values. Figure 4b shows a seemingly stable dependence plot that nevertheless misrepresents the prediction change. This discrepancy is caused by other features, like exercise-induced angina, whose SHAP values may also depend on age, as shown in Figure 4c. For additive models, shown in Figure 4d, SHAP values are functions of the corresponding feature and exactly recover the prediction change.

Local stability is not just an intuitive definition; it is also grounded in theory. As shown in Günther et al. (2025), SHAP values are not informative unless we have further assumptions about our model class. However, when SHAP values come with stability guarantees, they do become informative with respect to the definition in Günther et al. (2025) as the following theorem states.

Theorem 4 (Locally Stable SHAP Values Are Informative, Informal). *Consider the explanation $\mathcal{E}(f, x_0) = ((\Phi_1(x_0), \dots, \Phi_d(x_0)), U, (\varepsilon_1, \dots, \varepsilon_d))$, which, given a point x_0 , returns the SHAP values of x_0 , a neighborhood U around x_0 and tolerance thresholds $\varepsilon_1, \dots, \varepsilon_d$ such that the SHAP values are locally stable with respect to U and $\varepsilon_1, \dots, \varepsilon_d$. Then, $\mathcal{E}$ is an informative explanation if U has positive probability and the tolerance thresholds are sufficiently small.*

A formal version of the theorem, including exact bounds on ε_j as well as the proof, can be found in Appendix C.2 of the arXiv version. But even if SHAP values are locally stable, a SHAP dependence plot can still be misleading if our model is not additive.

Interpretation 2 and 3 Only Hold if the Model Is Additive. Interpretations 2 and 3 seem natural: If the plot shows an increase from age 43 to age 53, this might suggest that the risk of disease would be higher if our patient were older. Generally, the increase in the plot might raise the impression that the risk increases with age. However, Figure 4b shows a counterexample illustrating that neither of the two has to hold. There, we see the SHAP value of a patient in red as well as their nearest neighbors. Also here, it seems that increasing the patient’s age increases the prediction – but the opposite is true: The red curve shows the change in prediction if we vary the age of the patient represented by the red dot with all other features fixed, revealing that the prediction actually decreases with age. The reason for this is that the output is decomposed into the SHAP values by $f(x) = \Phi_1(x) + \dots + \Phi_d(x) + E(f(X))$ and the SHAP dependence plot only shows $\Phi_1(x)$. If the SHAP values of all

other features stayed the same when age is varied, this would imply that the SHAP plot exactly reflects the prediction behavior of the function (up to a constant). However, each SHAP value generally depends on all features, rather than just on its own feature, as has been pointed out repeatedly in the literature (Lundberg, Erion, and Lee 2018; Zhang et al. 2021; Bordt and von Luxburg 2023; Fumagalli et al. 2023). Hence, $\Phi_2, \dots, \Phi_d$ may also change when age is varied, making the prediction change in ways not reflected in the plot. For example, consider the feature of whether a patient has exercise-induced angina. One could imagine that the presence of such angina is only considered risky by the model if the patient is of a certain age. Figure 4c shows the SHAP values of the same patients as in Figure 4b, but this time for the feature “exercise induced angina”. And indeed, we observe that the SHAP values of other patients with exercise-induced angina vary from 0.3 to 0.5. Thus, this SHAP value is influenced by other features, possibly also by age. Similarly, a change in age could also cause a strong change in **all** other SHAP values, explaining why the red curve in Figure 4b shows such a different prediction behavior than what the SHAP plot suggests.

However, there is an assumption on the underlying model under which each SHAP value is invariant to change in other features: We call a function **additive** if it can be written as a sum of functions that only depend on one feature. That is, $f(x) = f_1(x^{(1)}) + \dots + f_d(x^{(d)})$ for component functions $f_1, \dots, f_d : \mathbb{R} \rightarrow \mathbb{R}$. To get rid of ambiguous constants in the components, one often separates the expected value and mean-centers the components, leading to $f(x) = f_1(x^{(1)}) + \dots + f_d(x^{(d)}) + E(f(X))$. Bordt and von Luxburg (2023) shows that in this case, every SHAP value is given by $\Phi_j(x) = f_j(x^{(j)})$. In particular, it only depends on the corresponding feature and is hence invariant to changes in other features. Reversely, if each SHAP value only depends on its own feature, the function f is additive as it is given by $f(x) = \Phi_1(x^{(1)}) + \dots + \Phi_d(x^{(d)}) + E(f(X))$. For an illustration, see Figure 4d. It shows a SHAP depen-

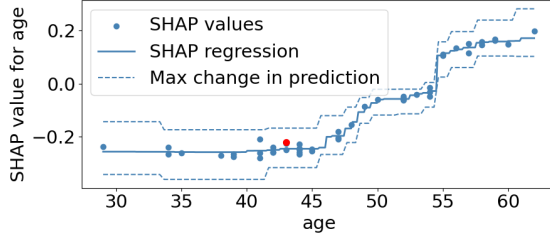
Explanation Card for SHAP Explanations in Medical Diagnosis, Targeting the Doctor

The patient is classified as sick with a predicted score of 0.565. In general, patients with positive scores are classified as sick; the higher the score, the more certain the classification algorithm.

The plots below show the SHAP values of patients in a local region around the explained instance:

Age:	29-62	Sex:	Male	Resting BP:	112-160 mmHG
Cholesterol:	157-325 mg/dl	Fasting blood sugar:	< 120 mg/dl	Max heart rate:	120-202
ST depression:	0-3	ST slope:	upsloping or flat		

Feature: Age



SHAP attributes a value of -0.22 to the patient's age (43), see red dot in the plot.

The SHAP values of other patients in the local region (blue dots) and their regression (blue line) roughly reflect how age affects the prediction. The dotted lines show the maximal/minimal change in prediction if age is varied.

Figure 5: Machine decision and SHAP value. Local region. Feature effect on prediction.

Explanation Card for a SHAP Explanation for a doctor debugging a prediction: In the explanation card, the SHAP explanations are augmented with human-intelligible interpretations, thereby helping prevent misinterpretation. They are split into general information, the prediction and a local region, and additional feature-wise plots. Depending on the model and observation, the card may also reveal that SHAP dependence plot does not reflect the prediction behavior as the dotted envelope is too large.

dence plot for an additive model, and again, a red curve illustrating the change in prediction as the age of the person represented by the red dot is varied. Unlike before, we now see that all SHAP values lie on a line (and are hence invariant to changes in other features) and precisely reflect the prediction behavior. This means that in the case of additive models, SHAP values do exactly what one would like them to do: They split the function value into a sum of feature contributions, which are completely determined by the feature itself. In particular, the SHAP dependence plot also describes the change in prediction with respect to that feature, rather than just the SHAP value.

In practice, it is unrealistic to expect our function to be additive. Nevertheless, for the doctor, it might be sufficient to understand the predicting behavior only in a local region, meaning for a given patient and similar patients. Also, we can further relax the problem by considering *approximate* statements. If in a local region, all SHAP values are mostly invariant to change in other features, Interpretation 2 and 3 will be approximately valid in that region. This leads to the following definition.

Definition 5 (SHAP Value Invariant to Change in Other Features). We say that the j -th SHAP value of a function f is approximately invariant to change in other features in a local region $U \subseteq \mathcal{X}$ and with respect to some error $\varepsilon > 0$ if

the expected unexplained variance of the SHAP value given its feature is less than ε :

$$\mathbb{E}_{X \sim U} \left(\text{Var}_{X \sim U} \left(\Phi_j(X) \mid X^{(j)} \right) \right) \leq \varepsilon.$$

The unexplained variance given feature $X^{(j)}$ measures how much variance there is left when we control feature j , and thus, how much the SHAP value varies due to a change in other features. With the aid of those quantities, one can now mathematically quantify to what degree a given SHAP value is invariant to change in the other features. As it turns out, this implies approximate additivity of the model in that local region, analogously to the global and exact case.

Theorem 6 (Local Approximate Invariance of SHAP Values Implies Local Approximate Additivity). *If every SHAP value of a function f is approximately invariant to change in other features with respect to some local region U and some error $\varepsilon > 0$, f can be approximated by an additive model $g(x) = g_1(x^{(1)}) + \dots + g_d(x^{(d)}) + \mathbb{E}(f(X))$ on U with mean squared error at most*

$$\mathbb{E}_{X \sim U} \left((f(X) - g(X))^2 \right) \leq d^2 \varepsilon.$$

The components of g are given by $g_j(x^{(j)}) = \mathbb{E}(\Phi_j(X) \mid X^{(j)} = x^{(j)})$.

The proof is in Appendix C.3 of the arXiv version. Although most models might not be globally additive, they might still yield an approximately additive structure in local regions, which would make an approximate, local version of Interpretation 2 and 3 realistic. The additive model g , as defined in Theorem 6, can be estimated by regressing the SHAP values with a one-dimensional model. The resulting mean-squared error can serve as a measure for how close the model is to being additive in the region of interest, and hence, how strongly a SHAP dependence plot links to the prediction there. This allows for choosing a region accordingly before further studying the SHAP values with our explanation card.

5.3 Explanation Cards Provide the User with Valid Interpretations

In this section, we propose an explanation card for SHAP, targeted towards a doctor aiming to understand a prediction. The explanation card is presented in Figure 5. It reports the SHAP values along with the information that the user needs to arrive at correct interpretations. In the header, the explanation card reports a local region around the explained instance in which the subsequent analyses are performed. Here, we consider the hyperbox defined by the 50 nearest neighbors with respect to the maximum-norm. Below, the card shows the SHAP values of patients within this region in a dependence plot, separately for each feature. Due to space constraints, we only present this for the feature “age”. For a full card, see Figure 8 in Appendix D of the arXiv version. In addition to the SHAP dependence plot, we display a regression of the SHAP values by the corresponding feature as an estimate for the effect of the that feature on the prediction. Also, for every point in the region, we vary the age and investigate the resulting change in prediction. We then take the maximum and minimum of the change in prediction over all points in the region, leading to the dotted lines in the plot. That means that if a feature value is replaced, the change in prediction stays within the two lines. They thus give an error bound for Interpretation 2 and 3. If the envelope between the two lines is too large (see Figure 6), the plot cannot be related to the prediction behavior and interpretations like 2 and 3 become invalid. For more technical details on how the card is derived from the SHAP explanations, see Appendix E.2 of the arXiv version.

In the example presented in Figure 5, the explanation card confirms an approximately additive structure with a small envelope, admitting approximate versions of Interpretation 2 and 3. In cases where they do not hold, like in Figure 4b, the explanation card reveals this: Figure 6 shows the same points as in Figure 4b but with the dotted lines, which form a big envelope. Another negative example of a SHAP explanation card with a huge dotted envelope is shown in Figure 9 in Appendix D of the arXiv version. It is important to highlight that an explanation card is not guaranteed to make sense of every explanation. Instead, it can tell us, for a given explanation, *whether* one can derive a specific intuitive interpretation from it, and if so, it can translate the mathematical information to human-understandable language. Especially in the case of SHAP dependence plots, complex models of

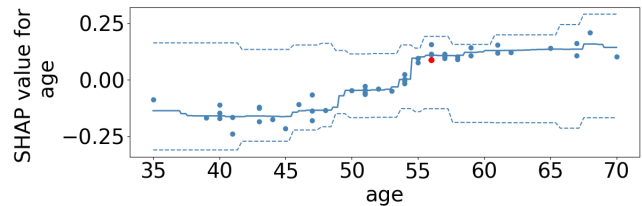


Figure 6: The fact that the dotted envelope is so large reveals that the SHAP dependence plot (blue points) does NOT reflect how the prediction function changes with age.

ten carry interactions, making it impossible to connect the plot to the prediction. The main point of this section and our card is to make these cases transparent, rather than to suggest a recipe for making SHAP values universally meaningful.

6 Explanation Cards and Transparency in the AI Act

General Setup of the EU AI Act. Within the European Union, the AI Act (AIA, Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744) sets specific rules for the use of AI and distinguishes providers, who develop an AI system, or have one developed, and place it on the market or put it into service under their own name or trademark (Art. 3(3) AIA), from deployers, who use a system under their authority in a professional capacity (Art. 3(4) AIA). Both definitions are simplified; see Kaminski and Selbst (2025) for an introduction to the AIA. Particular attention is given to Systems that are classified as high-risk (Art. 6 AIA). These must meet the requirements of Arts. 8 to 15 AIA, which aim to protect health, safety and fundamental rights (Recital 66 AIA; in detail Paul (2024)). Those requirements apply from 2 December 2027 to stand-alone Annex III systems and from 2 August 2028 to systems embedded in Annex I products (Art. 113 AIA as amended).

Transparency Requirements of Providers. Art. 13 AIA governs transparency from provider to deployer, with the aim of mitigating the risks of high-risk systems (Recitals 66 and 72 AIA) and providing usable explanations (Nanini, Balayn, and Smith 2023). High-risk systems must be “sufficiently transparent to enable deployers to interpret a system’s output and use it appropriately” (Art. 13(1) AIA); providers are bound to this via Art. 16, point (a) AIA. The instructions for use must therefore state, where applicable, the technical capabilities and characteristics relevant to explaining the output (Art. 13(3)(b)(iv) AIA) and the information deployers need to interpret that output and use it appropriately (Art. 13(3)(b)(vii) AIA). Throughout, the focus is the deployer’s understanding of the system’s capabilities and the limitations of its performance (Art. 13(3)(b), Recital 72 AIA). Instructions for use are a familiar transparency instrument in EU product safety law under the New Legislative Framework (Ortigosa 2025).

Explanation Cards could help address AI Act requirements. In practice, providers will rely heavily on harmonised standards. Currently being developed by CEN-

CENELEC, these standards will guide practical implementation (see amendment (Commission 2025), Kilian, Jäck, and Ebel (2025)); once their references are published in the Official Journal of the European Union, following them will establish a presumption of conformity (Art. 40(1) AIA). From our perspective, Explanation Cards could help address the requirements outlined in Article 13 AIA. As presented in Section 5, Explanation Cards offer multiple benefits in the context of Art. 13 AIA. They provide a clear, accessible, and comprehensive way to communicate explanations about the system’s outputs, focusing on the individual decisions made by the system. As a result, Explanation Cards help deployers understand the capabilities of the system. Additionally, Explanation Cards also highlight the limitations of the explanations themselves, offering a more nuanced perspective. This, in turn, leads to a better understanding when assessing the system’s performance and its limitations, as deployers gain greater insight into the entirety of the system.

Transparency Requirements Towards Individuals Affected by AI. Beyond the transparency requirements in Art. 13 AIA, Explanation Cards could be helpful for other legal transparency requirements to affected persons (See Section 4). While under Art. 13 AIA it is the explanation capability of the card that is relevant, the card could also serve as an instance-level artifact for an affected person. For example, they may help deployers fulfill the duty under Art. 86 AIA, which entitles a person affected by a decision taken by the deployer on the basis of the output of an Annex III high-risk system (excluding point 2 thereof), and producing legal effects or a similarly significant impact that the person considers adverse to their health, safety or fundamental rights, to clear and meaningful explanations of the role of the system in the decision-making procedure and of the decision’s main elements. Since that duty falls on the deployer and applies only where the right does not already follow from other Union law (Art. 86(3) AIA), Explanation Cards provided under Art. 13 AIA could serve as an input from which the deployer builds the explanation to the end user. Likewise, solely automated decision-making under Art. 22 in conjunction with Art. 15(1)(h) GDPR requires some form of explanation to the data subject (CJEU Case C-203/22 *Dun & Bradstreet* ECLI:EU:C:2025:117). On the relationship of these different “affected-person” explainability requirements, see Nannini (2024); Kaminski and Malgieri (2025).

Presenting an output together with its explanation card in a concise and understandable format can support clear communication of both the explanation and its limitations to professionals and layperson.

7 Discussion

Explanation Cards to Communicate Valid Interpretations. In this work, we propose Explanation Cards for Explanation Algorithms, an approach that makes explicit which interpretations and conclusions can be drawn from particular post-hoc explanations — and which ones cannot. The goal of the card is to connect the mathematically abstract concept of local post-hoc explanations to real-world applications and

to avoid misinterpretations of explanations. In order to derive reliable interpretations, explanation providers need to run additional analyses (for example, evaluate the function’s or explanation’s behavior in a neighborhood of the point of interest). In making these explicit, explanation cards provide additional information that goes beyond the standard output of particular explanation algorithms. Their generation can be automated by using text templates and simply filling in the computed values of the concrete application.

Making the Limitations of Interpretations Explicit. If an explanation does not admit the interpretations that humans tend to make, it is important to point this out to prevent false conclusions. For instance, the region of stability around a counterfactual explanation might be tiny, in which case it would be hard for an affected individual to change their features such that they are within this region (as in the right panel in Figure 1). Or the dotted envelope around the SHAP dependence plot might be so large that the plot cannot properly connect to the prediction (see Figure 6).

Concrete Examples: Counterfactuals and SHAP. Unlike related approaches, for example, datasheets for datasets (Gebu et al. 2021), there is no single explanation card that works for all explanation algorithms. The card needs to be designed for a particular explanation algorithm and a particular use case, taking into account the goals, desired interpretations, and background knowledge of the explanation receiver. We present two exemplary explanation cards: counterfactual explanations for loan decisions and SHAP explanations for medical diagnosis. Compared to SHAP, Counterfactuals have a very intuitive definition and meaning, causing the explanation cards to be easy to communicate to lay persons. For SHAP, on the other hand, it is less clear what the cards should contain, and the content naturally becomes more technical. We are curious whether simpler cards will be developed in future research.

Related Approaches. The regions of stability around the input point from Section 4.2 are similar in spirit to anchors (Ribeiro, Singh, and Guestrin 2018). We explicitly acknowledge this connection: In our view, counterfactual explanations should provide anchors that describe their region of validity. In the case of SHAP, the dotted lines are constructed using shifted ICE curves. A similar line of work suggests finding regions of the input space that are free of interactions (Gkolemis et al. 2024; Herbringer et al. 2024). These methods may potentially be used in implementations of the SHAP explanation card to find a suitable local region.

Legal Considerations. Explanations are especially important when it comes to high-stakes applications, particularly in societal contexts. The use of AI Systems for some of these decisions is regulated through the EU AI Act, which lays out requirements regarding transparency and explainability. We believe that Explanation Cards might provide a path forward to satisfy these requirements.

Future Work. In future work, intuitive interpretations that users assign to explanations could be studied systematically via user studies, together with the ability of Explanation Cards to reduce misinterpretations. Also, Explanation Cards for further explanation algorithms and use cases are left to be developed.

Acknowledgements

This work has been supported by the German Research Foundation through the Cluster of Excellence “Machine Learning - New Perspectives for Science” (EXC 2064/1 number 390727645) and Project 560788681, the Carl Zeiss Foundation through the CZS Center for AI and Law, and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Balázs Szabados was supported by the “Robust Uncertainty Quantification for Learning and Control” ADVANCED research project of the National Research, Development and Innovation Office of Hungary (NKFIH), grant number 153390.

References

- Adebayo, J.; Muelly, M.; Liccardi, I.; and Kim, B. 2020. Debugging Tests for Model Explanations. In *Neural Information Processing Systems (NeurIPS)*.
- Agarwal, A.; Lohia, P.; Nagar, S.; Dey, K.; and Saha, D. 2018. Automated test generation to detect individual discrimination in AI models. *arXiv preprint arXiv:1809.03260*.
- Artelt, A.; Vaquet, V.; Velioglu, R.; Hinder, F.; Brinkrolf, J.; Schilling, M.; and Hammer, B. 2021. Evaluating robustness of counterfactual explanations. In *IEEE symposium series on computational intelligence (SSCI)*.
- Bhatt, U.; Weller, A.; and Moura, J. M. 2021. Evaluating and aggregating feature-based model explanations. In *International Conference on Artificial Intelligence (IJCAI)*.
- Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M. F.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Boggust, A.; Suresh, H.; Strobel, H.; Guttig, J.; and Satyanarayan, A. 2023. Saliency cards: a framework to characterize and compare saliency methods. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Bordt, S.; Finck, M.; Raidl, E.; and von Luxburg, U. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Bordt, S.; Raidl, E.; and von Luxburg, U. 2025. Rethinking Explainable Machine Learning as Applied Statistics. In *International Conference on Machine Learning (ICML)*.
- Bordt, S.; and von Luxburg, U. 2023. From Shapley Values to Generalized Additive Models and back. In *Artificial Intelligence and Statistics (AISTATS)*.
- Borys, K.; Schmitt, Y. A.; Nauta, M.; Seifert, C.; Krämer, N.; Friedrich, C. M.; and Nensa, F. 2023. Explainable AI in medical imaging: An overview for clinical practitioners—Saliency-based XAI approaches. *European journal of radiology*, 162: 110787.
- Byrne, R. M. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *International Conference on Artificial Intelligence (IJCAI)*.
- Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.
- Comission, E. 2025. C(2025)3871 – Standardisation request M/613.
- Corbett, E.; and Denton, R. 2023. Interrogating the T in FAccT. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Dasgupta, S.; Frost, N.; and Moshkovitz, M. 2022. Framework for Evaluating Faithfulness of Local Explanations. In *International Conference on Machine Learning (ICML)*.
- Deck, L.; Schoeffer, J.; De-Arteaga, M.; and Kühl, N. 2024. A critical survey on fairness benefits of explainable AI. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Dhurandhar, A.; Haldar, S.; Wei, D.; and Natesan Ramamurthy, K. 2024. Trust Regions for Explanations via Black-Box Probabilistic Certification. In *International Conference on Machine Learning (ICML)*.
- Dombrowski, A.; Alber, M.; Anders, C. J.; Ackermann, M.; Müller, K.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In *Neural Information Processing Systems (NeurIPS)*.
- Dominguez-Olmedo, R.; Karimi, A. H.; and Schölkopf, B. 2022. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning (ICML)*.
- Ebers, M. 2020. Regulating explainable AI in the European Union. An overview of the current legal framework (s). *An Overview of the Current Legal Framework (s)(August 9, 2021)*. Liane Colonna/Stamley Greenstein (eds.), *Nordic Yearbook of Law and Informatics*.
- Fernández, R. R.; De Diego, I. M.; Aceña, V.; Fernández-Isabel, A.; and Moguerza, J. M. 2020. Random forest explainability using counterfactual sets. *Information Fusion*, 63: 196–207.
- Fokkema, H.; De Heide, R.; and Van Erven, T. 2023. Attribution-based explanations that provide recourse cannot be robust. *Journal of Machine Learning Research*, 24(360): 1–37.
- Freiesleben, T.; and König, G. 2023. Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In *World conference on explainable artificial intelligence*, 48–65.
- Freiesleben, T.; König, G.; Molnar, C.; and Tejero-Cantero, A. 2024. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *Minds and Machines*, 34(3): 32.
- Frye, C.; de Mijolla, D.; Begley, T.; Cowton, L.; Stanley, M.; and Feige, I. 2020. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*.
- Fumagalli, F.; Muschalik, M.; Kolpaczki, P.; Hüllermeier, E.; and Hammer, B. 2023. SHAP-IQ: Unified approximation of any-order shapley interactions.
- Garreau, D.; and von Luxburg, U. 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. In *Artificial Intelligence and Statistics (AISTATS)*.

- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Geirhos, R.; Zimmermann, R. S.; Bilodeau, B.; Brendel, W.; and Kim, B. 2024. Don't trust your eyes: on the (un)reliability of feature visualizations. In *International Conference on Machine Learning (ICML)*.
- Ghassemi, M.; Oakden-Rayner, L.; and Beam, A. L. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11): e745–e750.
- Gkolemis, V.; Diou, C.; Ntoutsis, E.; Dalamagas, T.; Bischl, B.; Herbringer, J.; and Casalicchio, G. 2024. Effector: A Python package for regional explanations. *arXiv preprint arXiv:2404.02629*.
- Grömping, U. 2019. South German Credit. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5X89F>.
- Günther, E.; Szabados, B.; Bhattacharjee, R.; Bordt, S.; and von Luxburg, U. 2025. Informative Post-Hoc Explanations Only Exist for Simple Functions. *arXiv preprint arXiv:2508.11441*.
- Hacker, P.; and Passoth, J.-H. 2020. Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. In *International workshop on extending explainable AI beyond deep models and classifiers*, 343–373. Springer.
- Hamon, R.; Junklewitz, H.; Sanchez, I.; Maltieri, G.; and De Hert, P. 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1): 72–85.
- Herbringer, J.; Wright, M. N.; Nagler, T.; Bischl, B.; and Casalicchio, G. 2024. Decomposing global feature effects based on feature interactions. *Journal of Machine Learning Research*, 25(381): 1–65.
- Huang, X.; and Marques-Silva, J. 2023. The Inadequacy of Shapley Values for Explainability. *arxiv preprint 2302.08160*.
- Hullman, J.; Guo, Z.; and Ustun, B. 2025. Explanations are a means to an end. *arxiv preprint arxiv:2506.22740*.
- Hwang, H.; Bell, A.; Fonseca, J.; Pliatsika, V.; Stoyanovich, J.; and Whang, S. E. 2025. SHAP-based Explanations are Sensitive to Feature Representation. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Janzing, D.; Minorics, L.; and Blöbaum, P. 2020. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on artificial intelligence and statistics (AISTATS)*.
- Kaminski, M. E.; and Maltieri, G. 2025. The Right to Explanation in the AI Act. *U of Colorado Law Legal Studies Research Paper No. 25-9, available at SSRN 5194301*.
- Kaminski, M. E.; and Selbst, A. D. 2025. An American's Guide to the EU AI Act. *Available at SSRN 5373345*.
- Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5): 1–29.
- Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Khadar, M.; Timma, A.; Ludden, L.; and Kaur, H. 2025. Explain it like I'm me: User Characteristics in XAI Decision-Making. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Kilian, R.; Jäck, L.; and Ebel, D. 2025. European ai standards—technical standardisation and implementation challenges under the eu ai act. *European Journal of Risk Regulation*, 1–25.
- König, G.; Fokkema, H.; Freiesleben, T.; Mendl-Dünner, C.; and von Luxburg, U. 2025. Performative validity of recourse explanations. *arXiv preprint arXiv:2506.15366*.
- König, G.; Freiesleben, T.; and Grosse-Wentrup, M. 2023. Improvement-focused causal recourse (ICR). In *AAAI Conference on Artificial Intelligence*.
- König, G.; Günther, E.; and von Luxburg, U. 2025. Disentangling Interactions and Dependencies in Feature Attributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Krishna, S.; Han, T.; Gu, A.; Pombra, J.; Jabbari, S.; Wu, S.; and Lakkaraju, H. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602*.
- Krishnan, M. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3): 487–502.
- Kumar, I. E.; Venkatasubramanian, S.; Scheidegger, C.; and Friedler, S. A. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning (ICML)*.
- Laugel, T.; Lesot, M.-J.; Marsala, C.; and Detyniecki, M. 2019. Issues with post-hoc counterfactual explanations: a discussion. In *ICML Workshop on Human in the Loop Learning (HILL 2019)*.
- Leofante, F.; and Potyka, N. 2024. Promoting counterfactual robustness through diversity. In *AAAI Conference on Artificial Intelligence*.
- Liao, Q. V.; and Varshney, K. R. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.
- Lipton, Z. C. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.
- Lundberg, S. M.; Erion, G. G.; and Lee, S.-I. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.
- Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Neural Information Processing Systems (NeurIPS)*.
- Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.
- Molnar, C. 2025. *Interpretable Machine Learning*. 3 edition.
- Molnar, C.; König, G.; Herbringer, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C. A.; Casalicchio, G.; Grosse-Wentrup, M.; and Bischl, B. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*.
- Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- Nannini, L. 2024. Habemus a Right to an Explanation: so What?—A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act. In *Conference on AI, ethics, and society (AIES)*.
- Nannini, L.; Balayn, A.; and Smith, A. L. 2023. Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Conference on fairness, accountability, and transparency (FACCT)*.
- Ortigosa, A. P. 2025. Transparency and Provision of Information to Deployers (Article 13). In *The EU regulation on Artificial Intelligence: A commentary*, 217–242. Wolters Kluwers Italia.
- Panigutti, C.; Hamon, R.; Hupont, I.; Fernandez Llorca, D.; Fano Yela, D.; Junklewitz, H.; Scalzo, S.; Mazzini, G.; Sanchez, I.; Soler Garrido, J.; and Gomez, E. 2023. The role of explainable AI in the context of the AI Act. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- Paul, R. 2024. European artificial intelligence “trusted throughout the world”: Risk-based regulation and the fashioning of a competitive common AI market. *Regulation & Governance*, 18(4): 1065–1082.
- Raees, M.; Meijerink, I.; Lykourantzou, I.; Khan, V.-J.; and Papangelis, K. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies*, 189: 103301.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI Conference on Artificial Intelligence*.
- 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.
- Schmitt, V.; Villa-Arenas, L.-F.; Feldhus, N.; Meyer, J.; Spang, R. P.; and Möller, S. 2024. The role of explainability in collaborative human-AI disinformation detection. In *Conference on fairness, accountability, and transparency (FACCT)*.
- Schor, B. G. S.; Kallina, E.; Singh, J.; and Blackwell, A. 2024. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- Selbst, A.; and Powles, J. 2018. “Meaningful information” and the right to explanation. In *Conference on Fairness, Accountability and Transparency (FACCT)*.
- Skirzynski, J.; Danks, D.; and Ustun, B. 2025. Discrimination Exposed? On the Reliability of Explanations for Discrimination Detection. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- Slack, D.; Hilgard, A.; Lakkaraju, H.; and Singh, S. 2021. Counterfactual explanations can be manipulated. In *Neural Information Processing Systems (NeurIPS)*.
- Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Conference on AI, Ethics, and Society (AIES)*.
- Taufiq, M. F.; Blöbaum, P.; and Minorics, L. 2023. Manifold restricted interventional shapley values. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Tomaszewska, P.; and Biecek, P. 2024. Position: Do Not Explain Vision Models Without Context. In *International Conference on Machine Learning (ICML)*.
- Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- VanNostrand, P. M.; Hofmann, D. M.; Ma, L.; and Rundensteiner, E. A. 2024. Actionable recourse for automated decisions: examining the effects of counterfactual explanation type and presentation on lay user understanding. In *Conference on Fairness, Accountability, and Transparency (FACCT)*.
- VanNostrand, P. M.; Zhang, H.; Hofmann, D. M.; and Rundensteiner, E. A. 2023. Facet: Robust counterfactual explanation analytics. *Proceedings of the ACM on Management of Data*, 1(4): 1–27.
- Velmurugan, M.; Ouyang, C.; Moreira, C.; and Sindhgatta, R. 2021. Evaluating stability of post-hoc explanations for business process predictions. In *International Conference on Service-Oriented Computing*.
- Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.; Dickerson, J.; and Shah, C. 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12): 1–42.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841.
- Zhang, H.; Xie, Y.; Zheng, L.; Zhang, D.; and Zhang, Q. 2021. Interpreting multivariate shapley interactions in dnn. In *AAAI Conference on Artificial Intelligence*.

Zhang, S.; Chen, X.; Wen, S.; and Li, Z. 2023. Density-based reliable and robust explainer for counterfactual explanation. *Expert Systems with Applications*, 226: 120214.