SIDEs: Separating Idealization from Deceptive ‘Explanations’ in xAI

Emily Sullivan
e.e.sullivan@uu.nl
Utrecht University
Utrecht, Netherlands

ABSTRACT

Explainable AI (xAI) methods are important for establishing trust in using black-box models. However, recent criticism has mounted against current xAI methods that they disagree, are necessarily false, and can be manipulated, which has started to undermine the deployment of black-box models. Rudin (2019) goes so far as to say that we should stop using black-box models altogether in high-stakes cases because xAI explanations ‘must be wrong’. However, strict fidelity to the truth is historically not a desideratum in science. Idealizations–the intentional distortions introduced to scientific theories and models–are commonplace in the natural sciences and are seen as a successful scientific tool. Thus, it is not falsehood *qua* falsehood that is the issue. In this paper, I outline the need for xAI research to engage in *idealization evaluation*. Drawing on the use of idealizations in the natural sciences and philosophy of science, I introduce a novel framework for evaluating whether xAI methods engage in successful idealizations or deceptive explanations (SIDEs). SIDEs evaluates whether the *limitations* of xAI methods, and the distortions that they introduce, can be part of a successful idealization or are indeed deceptive distortions as critics suggest. I discuss the role that existing research can play in idealization evaluation and where innovation is necessary. Through a qualitative analysis we find that leading feature importance methods and counterfactual explanations are subject to idealization failure and suggest remedies for ameliorating idealization failure.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Social and professional topics → Technology audits**.

KEYWORDS

explainable AI, idealization, philosophy of science, qualitative evaluation

ACM Reference Format:
Emily Sullivan. 2024. SIDEs: Separating Idealization from Deceptive ‘Explanations’ in xAI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658999

FAccT ’24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658999

1 INTRODUCTION

The ideal gas law dating back to 1834, along with its simpler cousin Boyle’s law from the 1660s, are still used to explain how gases behave. Boyle’s law captures the inverse relationship between pressure and volume at constant temperatures, which can explain why a balloon’s inflation level changes amidst elevation changes. The ideal gas law complicates the picture by adding the influence of temperature changes and molarity to gas behavior. However, both these laws involve distortions and departures from the truth [27, 78]. Real gases do not behave ideally. Particles are assumed not to interact (even though they do), and the actual relationship between pressure and volume is more complicated than either law lets on. Despite this, the ideal gas law is still highly successful–a highly successful idealization (i.e. an intentional distortion introduced to scientific theories and models).

Contrast this with a more controversial contemporary case. In 2021, RIVM, the Dutch health institute, constructed a model to measure the spread of nitrogen pollution. In order to reduce computational complexity, the model treated nitrogen deposits coming from roads as spreading only 5km, while the deposits from farms traveled much longer distances. However, this distortion–this idealization–was not a success, and was later removed because it disproportionately put the pollution blame on farms instead of highways [104]. What makes the ideal gas law a permissible, even desirable, distortion, but the idealizations in the Dutch nitrogen model problematic, needing revision? It cannot simply be that the Dutch nitrogen model contains falsehoods *vis a vis* an idealization, or a lack of fidelity to the phenomena, since the ideal gas law does the same.

Now consider the topic at hand: explainable AI (xAI) methods. Rudin [85] argues that the increasing trend of using black-box ML models across science and society is problematic, precisely because the methods we use to interpret these models provide us with necessarily wrong explanations of how they work. Other recent works have painted a picture of mounting criticism that leading xAI techniques are unreliable, subject to manipulation, and different techniques often disagree [2, 21, 48, 91, 92]. Some critics argue that xAI is a ‘false hope’ [31], where methods can engage in explanation hacking [99] or fairwashing [1], leading users to unduly trust models [32, 51, 59], and more [30]. This raises the question: Do the falsehoods and approximations (i.e. idealizations) operating in xAI have the same secret of success as the ideal gas law? Or are the critics right that current xAI methods need improvement before systematic deployment, like the Dutch nitrogen case? When are deviations from the truth successful idealizations? When are they deceptive distortions?

This paper seeks to animate a new research program for the study of xAI *idealizations*. Idealization evaluation moves beyond concepts of falsehood or model fidelity capturing a deeper issue: the norms and practices surrounding model limitations and their context of appropriate use. I introduce a broad evaluation framework for idealization evaluation in xAI that aims to separate successful idealizations from deceptive explanations (SIDEs) (sect. 4). I identify areas where existing xAI evaluation methods are useful for idealization evaluation, and where innovation is necessary. To build this framework, I first take inspiration from the idealization practices in the natural sciences (sect. 3). I take an off-the-shelf theory of idealization influential in philosophy of science, the minimalist view of idealization, and apply it to feature importance (sect. 4) and counterfactual explanation methods (sect. 5), showing through a qualitative analysis that idealization failure is common. Lastly, I consider ways of ameliorating idealization failure and point to future research directions for building a theory of idealization for xAI (sect. 6). It is also important to say what this paper does *not* do. It does not argue for a new philosophical theory of idealization or seek to diagnosis exactly what kind of idealization xAI methods engage in. Instead, I provide a modular framework for and show the potential for xAI researchers to engage with work on idealization in philosophy of science. The paper aims to make the following contributions:

(1) Conceptualizes the field of xAI as solving an idealization problem
(2) Introduces the SIDEs framework for evaluating idealizations in xAI, derived from normative foundations of idealization in the natural sciences and philosophy of science.
(3) Provides a theoretical grounding for novel evaluation methods for xAI.
(4) Identifies possible novel practices of idealization in xAI that needs normative analysis.

2 BACKGROUND AND RELATED WORK

The ever-growing fingerprint ML has on the production of scientific and social knowledge comes with challenges. One often cited issue is model opacity [12, 16, 20]. Transparency of ML decisions is important for building trust [35, 41, 59, 63], it might be legally required [34, 88], and convincing arguments have been made for a moral right to explanation in high-impact contexts [106]. This need for transparency inspired a proliferation of different interpretability and xAI techniques to solve the problem of opacity by providing insight into the reasons behind ML classifications. Post-hoc feature importance methods remain the most influential approach in xAI. These methods seek to approximate how much particular features contribute to the model's decision locally around each prediction. Examples include LIME [81], SHAP which utilizes coalitions game theory [61], and saliency maps that visualize regions of interest [3, 90]. These methods differ from example-based or decision-rules [67], and differ from global explanation methods that seek to capture the behavior of a black-box system as a whole [40, 52]. However, there are ways to use LIME and SHAP to get close to a global explanation by aggregating many local explanations. Counterfactual explanation (CE) methods have recently gained notoriety as the leading alternative to feature importance methods [5, 37, 62, 68, 86, 103, 105, 107]. CE methods seek to answer *what-if-things-had-been-different* questions by probing the ML model to see what minimal changes would reverse the ML decision. There are a variety of different algorithms for generating or filtering which counterfactuals would be relevant in different contexts and for different stakeholders (see [42] for a review). Despite these acheivements there remain conceptual and evaluative challenges to xAI and explainability.

Conceptual issues for xAI. There are several conceptual contributions that philosophers of science have made in debates around xAI. Most notably, on the concept of *explanation* [63]. Central questions concern what type of information is required to fulfil the definition of an 'explanation' in philosophy of science [15, 28, 63, 70, 72], such as conforming to a covering-law view of explanation [28] or having an additional link between the model and the world [96]. Others focus on ethical considerations regarding whether certain xAI methods can fulfil moral requirements for explanation [105, 106], such as a principled reason explanation [5]. These normative based approaches have exposed various challenges with xAI. For example, Symons and Alvarado [102] discuss the issue of epistemic injustice in the context of trying to solve ML opacity. Others have suggested that CE methods have the potential to hide bias [1, 5, 100, 101]. While even some argue that xAI methods are unnecessary and that model evaluation should focus instead on notions of reliability [24, 25, 36, 60]. Discussions of the norms of *explanation* are no doubt important and necessary. However, as I hope to show in this paper, norms of explanation are distinct from the norms and ideals that govern idealization. However, only recently has their been a suggestion that the concept of idealization in xAI may be useful [10, 29]. Nevertheless, here researchers stop short of discussing how xAI researches could actually use idealization theory, and how we could evaluate idealizations.

Current approaches to xAI evaluation. Evaluating xAI methods in computer science include experimental methods, such as comparisons of accuracy and model fidelity between different algorithms and benchmarks [7, 59], whether methods are robust under manipulation or perturbations [2, 21, 91, 92], or whether such methods conform to human expectations [33, 65, 69]. Current evaluation methods have exposed a number of vulnerabilities. Accuracy tests conducted on feature importance methods, found that the best performing method only approached 85% agreement with the black-box model, with LIME often scoring lower [52]. Furthermore, feature importance methods were found to be vulnerable to adversarial manipulation. Slack et al. [92] were able to create explanations that hid the most salient feature for classification for SHAP and LIME. Ghorbani et al. [32] found such methods were highly sensitive to small changes to input data, with others finding that they are not able to capture causal notions [59, 76, 85]. Counterfactual explanation methods can also fall prey to manipulation. Specifically, it was found that hill-climbing CE methods can converge to different local minima resulting in possible manipulation [91]. They also suffer from the rashomon effect where different counterfactuals explaining the same decision can be contradictory [17, 56].

While current methods of xAI evaluation are no doubt insightful, important gaps remain. Current experimental methods actively look for vulnerabilities and look for cases where methods break down or where new methods show an improvement compared to benchmarks. However, they stop short of providing an analysis that evaluates whether the *limitations* and the distortions xAI methods introduce are actually problematic or could be a case of a successful idealization. Instead, several critics simply point to the existence of possible manipulation and limited model fidelity, as itself a strike against the method [2, 48, 51, 59, 85]. While others have argued that current notions of fidelity are ill-equipped to capture cases of misleading explanation [7, 51]. The potential for misleading explanations has also inspired a user-centered approach to evaluation, where xAI evaluation is geared toward fulfilling either actual user preferences, or expected user perceptions of usefulness [41, 84, 109], including the introduction of normative stakeholder sensitive frameworks [22, 44, 54, 66, 114]. Again, while this evaluative approach is important, normative and theory-based evaluations concerning the potentially *positive* role a lack of model fidelity could have for xAI methods are lacking.

Closing Gaps. This paper aims to address the above gaps by introducing the concept and framework of idealization evaluation. While model auditing techniques look for vulnerabilities, idealization evaluation asks whether such vulnerabilities are problematic or actually a successful tool. Moreover, idealization evaluation provides the conceptual tools for identifying the fundamental goals for xAI more so than just looking at theories of explanation. With this paper, I hope to show the need for xAI research to build a theory of idealization and engage directly in idealization evaluation. Without a proper theory of idealization, it remains difficult to thoroughly diagnose the success of xAI methods. Idealization is inevitable, but if done right, idealization is desirable. My approach in this paper takes inspiration from idealization in the natural sciences and the philosophy of science to gain insight into how researchers can begin the project of idealization evaluation in xAI. I return to how idealization evaluation fits within current xAI research in section 4.

3 IDEALIZATION IN THE NATURAL SCIENCES AND PHILOSOPHY OF SCIENCE

Idealizations are the (intentional) distortion of real-world features that are present in a model or theory. In science idealizations are many. Examples include the ideal gas law and frictionless planes in physics, perfectly rational agents in economics, infinite populations and the absence of genetic drift in biology, etc. The way philosophers of science understand the concept of idealization might be best illustrated with an example outside of science. The Tube map of London's underground has neatly organized lines, and the circle line resembles a circle. However, the Tube map distorts the actual layout of the Tube tunnels. In reality, the interconnection of tunnels is complex and rarely a straight line [80]. The official Tube map does more than leave out detail; it intentionally distorts the real layout of tunnel paths. The Tube map idealizes London's subway structure. Philosophers of science have sought to understand and conceptualize the nature, function, and epistemic value of idealizations in scientific inquiry [27, 55, 100].

Table 1: Features of Idealization

Idealization Features	Description
Purpose	The purpose / function of the idealization (e.g. epistemic purpose, like understanding; ethical purpose, like recourse and contestability, etc.)
Idealization Practice	The set of scientific methods and practices that categorize a type of idealization, along with the justification of those practices (e.g. Minimalist idealization)
Ideals	Values and norms underlying an idealization practice that govern rule development (e.g. causal-entailment, user feasibility)
Rules	The way ideals are operationalized into a metric of evaluation
User-facing Explanations	How idealizations are presented as explanations to end-users

Features and types of idealization practices: In a landmark paper in philosophy of science, Weisberg [110] proposed that idealizations should be categorized by their specific scientific practices made up of the activity of scientists, the norms or values that govern these practices, and how these norms are justified. We can translate these aspects of idealization into three features (idealization practice, ideals, rules) that are important in the natural sciences, adding two additional features for xAI (See Table 1). Philosophers of science have conceptualized several different idealization practices in the natural sciences, like the unique quality of infinite idealizations [89] and asymptotic idealizations in physics [8, 95], hypothetical-pattern idealization in biology [83], the practice of Galilean idealization, multiple-model idealization [110], and more. In this paper, we restrict discussion to one influential theory of idealization, Strevens' [93, 94] minimalist view of idealization (MinI). Below I will discuss how MinI works in a simple physics case, and in the next section discuss why MinI resembles the idealization practices in xAI and is a good place to start for evaluating xAI's idealizations. In section 6, I discuss alternative idealization practices that xAI could be engaged in.

Minimalist idealization in physics. The underlying norm for MinI is that simple models and explanations are better than more complex ones. Understanding phenomena requires isolating relevance from irrelevance often requiring idealization. MinI's idealization practice consists of devising scientific (or mathematical) methods for reducing the number of features that give rise to a phenomenon, highlighting the *difference-makers*, and only distorting the *non*-difference-makers. As such, the governing ideals for MinI are simplicity and isolating difference-makers [93, 94]. In physics, relevance and difference-making is usually a type of causal or dependency entailment, where some causal consequence can be (logically) derived from a set of initial conditions along with a causal law. In cases where the law is non-causal, the entailment is a different type of dependency entailment (such as a mathematical dependency) [53]. While Strevens [93] focuses on causal difference-making, others have adopted non-causal approaches to MinI [9].

The most discussed example of MinI is the ideal gas law. The ideal gas law introduces the false assumption that a system consists of N non-interacting particles so that physicists can clearly see that phase space is proportional to volume. The justification for adding the idealization of non-interacting particles is that in contexts of low pressure and high temperature, particle interactions are virtually insignificant to the relations between pressure, temperature, molarity, and volume. The idealization highlights this irrelevancy in a way that a more accurate representation hides.

However, if we remove the idealization, we can still determine the irrelevance of particle interactions. We can derive the virial equation of state directly from statistical mechanics with arbitrary precision by extending the equation indefinitely, where each added term is derived from an increasingly detailed and accurate representation. However, the contribution of each added term becomes vanishingly small, again resulting in the ideal gas law [100]. Thus, the ideal gas law satisfies the inclusion and fidelity rules of MinI by only removing and distorting aspects that do *not* affect causal-entailment (i.e. only distorting non-difference-makers). In cases where entailment fails and the ideal gas law does not capture gas behavior (e.g. in high pressure), other laws are required (e.g. Van der Waals). Importantly, even if it is possible to de-idealize in this case, MinI is still appropriate. MinI captures the difference-makers that a more accurate alternative does not. As such, idealized distortions are permanent fixtures–even if they can in principle be removed–because they distinguish relevance from irrelevance.

Idealization in xAI?. Machine learning is not the type of practice that philosophers of science have built their idealization theories around. ML models are complex instead of simple and they are not constructed with built-in theoretical assumptions where model equations explicitly represent processes in the target system [46]. ML models are often used precisely because such theoretical assumptions are unavailable, or because researchers are interested in prediction or overlooked patterns of interest. Moreover, in the case of xAI, the xAI model is an idealization not of the world but another model (the ML model). Thus, we need to separate between two questions of xAI and ML idealizations:

- **Model-World question**: How do black-box ML models idealize some real-world phenomenon? (e.g. how do ML models idealize aspects of disease indicators?) [23, 96, 97]
- **Model-Model question:** How do xAI methods idealize how a black-box ML model works? How is an xAI method an idealized representation of the black-box model? (e.g. how do feature importance methods idealize aspects of the ML model decision process?)

XAI mainly concerns the Model-Model question (we return to Model-World questions for xAI in sect. 5). Like the ideal gas law, there are several similarities between the xAI project and MinI. Current work in xAI often describes the ultimate goal of xAI methods as uncovering how black-box models make decisions, capturing how various inputs can cause a particular output in the black-box model [63, 76, 108]. This leaves Fleisher [29] to argue that feature importance methods are a kind of MinI because they satisfy simplification, flag difference-makers, and focus on a specific causal pattern in the black-box model that gives rise to the decision (i.e. answering the model-model question). For example, LIME uses linear approximation methods that distort aspects of the black-box model decision making process, but does so by aiming to find the features that are the central difference-makers for a given local decision (e.g. high debt is the largest difference-maker for why the black-box model recommended loan rejection). But are feature importance successfully engaging in MinI, like the ideal gas law?

4 TOWARD A FRAMEWORK FOR IDEALIZATION EVALUATION IN XAI

In this section, I introduce the SIDEs framework. SIDEs consists of a high-level modular workflow (Figure 1) that can guide researchers with key questions for reflection and qualitative evaluation of xAI idealizations. I go through each phase of SIDEs, identifying areas where existing theories and experimental evaluation techniques are useful and where innovation is required. If idealizations meet the standards for each phase, with alignment between phases, then the idealization is successful. Idealization failure occurs when there is misalignment or the idealization fails to meet the standards for a given phase. Throughout this section, leading feature importance methods, LIME and SHAP, are qualitatively evaluated to illustrate how SIDEs can identify risks of idealization failure. Section 5 considers CE methods.

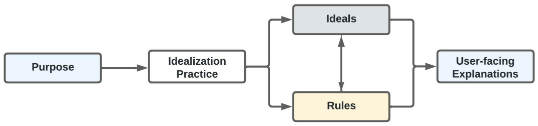

Figure 1: SIDEs Workflow

4.1 Purpose

In philosophy of science idealization analysis begins with an idealization practice. However, xAI calls for starting idealization analysis with the purposes that researchers are aiming to achieve with idealizations [14]. In the natural sciences, idealizations are discussed in the context of ideal scientific agents, so the central purpose of an idealization is presupposed to be epistemic by enabling understanding of phenomena [78, 82]. However, xAI methods serve a variety of purposes beyond epistemic purposes, including ethical purposes like recourse, where the aim is informing end-users about actionable changes that could reverse a negative decision [103]. To capture this difference, the SIDEs workflow starts with identifying the overall purpose the xAI method serves in a given context.

I want to highlight two potentially conflicting purposes of xAI methods: epistemic and ethical purposes. However, additional purposes are possible (e.g. legal compliance). While it can be helpful to discuss purposes of xAI broadly, finer grained aims are more useful for idealization evaluation. For example, fine-grained epistemic purposes could be understanding the phenomena the model bears on [96], predictive knowledge [38], user-specific epistemic goals [98], and more [13, 73]. Ethical purposes include fairness [4], providing users with recourse (i.e. actionable interventions) [105], exposing model bias, etc. Purposes are not necessarily mutually exclusive and may overlap. A single xAI explanation might aim to satisfy several purposes at once, such as providing users understanding of the model and building users' trust in the model, while also providing a user with algorithmic recourse. SIDEs does not preclude xAI methods from serving multiple purposes; however, the bar for successful idealization could become considerably higher (see section 5).

4.1.1 *Evaluating Purpose.*

The PURPOSE phase in SIDEs asks researchers to reflect on what purposes an xAI method does and *should* have in a specific context. There are several existing works that have identified various purposes that xAI methods serve [49, 59], and there is normative and theory-based work on xAI concerning what purposes xAI should have that can be helpful for evaluating PURPOSE [44]. A central pitfall for xAI methods is the risk of idealization failure due to misalignment with purpose. For example, in the original LIME paper, Ribeiro et al. [81] describe the motivating purpose behind LIME as establishing user trust. However, trust could have divergent underlying aims. On the one hand, trust can serve an ethical purpose. In clinical cases, patients trust of a doctor's diagnosis is often not grounded in the patient's knowledge or understanding of the diagnosis [64, 112]. On the other hand, there are other contexts where trust is only achieved when users understand the reasoning behind decisions. This tension between different functions of trust complicates the picture of whether an xAI method engages in idealizations that could fulfil these various purposes. For example, Lakkaraju and Bastani [51] found LIME is able to manipulate user trust.

4.1.2 *Role and limits of current work.*

Current research directions in xAI are well-placed to evaluate the purposes that xAI methods do and should have, with significant work already being done [14, 41, 49, 59, 70]. Re-conceptualizing xAI as an idealization problem relies on this work as the first fundamental step toward establishing how model fidelity should be understood and which features of black-box models can be distorted (i.e. idealized). Idealization evaluation asks researchers when analyzing purposes of xAI to consider the extent to which model fidelity matters.

Key questions for the PURPOSE phase are:

- *What purpose does an xAI explanation have in a particular context? What purpose should it have?*
- *What aspects of the model need to be known for a particular purpose?*
- ***Example**: What notion of trust is an explanation aiming for in a particular context?*
- ***Success**: The purpose of the xAI explanation is appropriate for the given deployment context.*

4.2 Idealization practices

As discussed in section 3, idealization practices consist of the set of scientific methods and practices along with the justification of those practices. One central research area in philosophy of science is conceptualizing different idealization practices across the sciences into distinguishable types or theories of idealization.

4.2.1 *Evaluating Idealization Practices.*

Evaluating the IDEALIZATION PRACTICES phase involves two central aspects. The first is a descriptive project that systematizes current work in xAI, elucidating a set of common aims and methodologies. This can be done for xAI in general or for a specific class of xAI methods. Second, evaluating IDEALIZATION PRACTICES consists of a justification step that can ground the legitimacy of the idealization practice. Using MinI as our working hypothesis, the methodology of MinI consists in omitting or distorting (causal) influences for the purposes of highlighting the central (causal) difference-makers or (causal) patterns. MinI is justified both through a strong conceptual foundations in scientific understanding, explanation, and (casual) difference-making, and in its empirical success [55].

4.2.2 *Role and limits of current work.*

Currently there has been very little work trying to conceptualize the type of idealization practices computer scientists are engaged in when developing xAI methods [29], and these practices are still arguably elusive [59]. However, as discussed in sect. 3, there are several similarities between current xAI methods and MinI. XAI aims to cut through the noise of many feature interactions to arrive at the chief difference-makers for a decision. For the purposes of this paper, we evaluate xAI methods as if they are engaging in the idealization practice of MinI. However, this paper calls for xAI to actively engage in solidifying one or more idealization practices and to work with philosophers of science on establishing conceptual foundations that justify these idealization practices. This is a central area that requires innovation and future research (see also section 6).

Key questions for the IDEALIZATION PRACTICES phase are:

- *What are the specific methods of deriving idealizations (e.g. introducing certain idealization assumptions, mathematical operations applied on data that results in distortions of the phenomena)?*
- *Does the idealization practice align with the purpose of the idealized model?*
- *What justifies this particular idealization practice? Why is it suitable for the identified purpose?*
- ***Example:** Minimalist idealization provides better understanding of the relevant difference-makers.*
- ***Success:** The idealization practice is well-grounded and justified in a specific domain, aligning with PURPOSE.*

4.3 Ideals and Rules

IDEALS are the norms and values that govern a specific idealization practice. RULES are the operationalization of these norms and values. For example, the ideals for MinI are to isolate the minimum number of difference-makers that capture a phenomenon, while Strevens [93] describes the fidelity and inclusion rules of MinI as satisfying a 'causal entailment' test, where modelers remove (or distort) features, finding the minimal amount that still entail the desired event.

4.3.1 *Evaluating IDEALS and RULES.*

Evaluating idealizations in the IDEALS and RULES phase requires validating whether specific rules embody target ideals, and analyzing trade-offs between different ways of operationalizing ideals. Idealizations can be experimentally validated by developing tests that satisfy these rules. Adopting our working hypothesis that xAI is engaging in MinI means that for xAI methods to be a legitimate idealization of a black-box model and pass the IDEALS and RULES phase, xAI methods must isolate the difference-makers for the target black-box model's decision by undergoing a (causal) entailment test that ensures the xAI method gives the same results as the target black-box model. Passing an entailment test would mean, in theory, that the xAI method uncovers the minimum set of difference-makers that determine the black-box model's decision [108], and thus any distortion it makes in the process is a legitimate one.

However, specifying an adequate (causal) entailment test is not simple. First, the notion of difference-making in xAI must be clear. For Strevens [93, 94], MinI aims for causal difference-making, where the notion of causality is left implicit and entailment more closely resembles logical deduction. Comparatively, philosophical theories like an interventionist framework [108, 113] or a counterfactual framework [45, 58] would result in grounding different causal rules. Alternatively, MinI need not aim for *causal* difference-making at all; other notions of difference-making are consistent with MinI (logical, probabilistic, mathematical, etc.). Thus, it is necessary to establish the specific ideals that a given idealization seeks to capture. Second, even once we settle on a notion of difference-making for MinI, there are still different possible rules that could capture the norms of MinI and serve as a basis for idealization evaluation. In this section, I consider three possible rules based on Strevens view of MinI. My aims are to 1) illustrate how ideals might be operationalized (see Table 2); 2) discuss how to think about trade-offs and whether a certain rule embodies the target ideal; and 3) discuss where existing evaluation methods in xAI are useful and areas where innovation is needed.

Table 2: Entailment RULES

MAP	ELIMINATION	PROB
$\forall x\ b(x) = e(x)$	$\forall Y \subset I_e \quad b(X - Y) = b(X)$	$\forall x\ P(\textsc{rule}(x)) > t$

First, MinI entailment could be a global entailment where the xAI model shares the same mapping of model inputs to outputs as the black-box model. The MAP rule requires that the mapping of inputs and outputs for an xAI model, *e*, is the same as for the black-box model, *b*. Map does not look at the features that the xAI model highlighted as relevant. As long as there is a 1 to 1 input-output mapping, then this is enough to establish a global notion of (causal) entailment. Map aligns with some of the current approaches to xAI evaluation. Accuracy and model fidelity metrics used for feature importance methods aim to see how well explanations mirror black-box predictions [51]. The results of these tests have exposed important vulnerabilities. For example, Lakkaraju et al. [52] found the best performing method only approached 85% agreement with the black-box model, with LIME often scoring lower. Even if we have a 1-1 mapping, an important trade-off to consider with MAP is how well it aligns with the purpose for xAI. Usually the purpose of an xAI method is for users to learn about the reasons for why the black-box model made its decision, not merely that an alternative proxy-model (in the case of LIME and SHAP) can derive the same predictions, which is why Lakkaraju et al. [52] argue that high-fidelity xAI models aren't enough.

Second, as an alternative, we could operationalize entailment more closely with Strevens' [93] own elimination test for MinI in the natural sciences. ELIMINATION tells us that for all the features *Y* that are in the set of putatively irrelevant features *I* found from the xAI model *e*, we can remove those features from the set of all input features *X* from the black-box model *b* and still receive the same decision. According to ELIMINATION, evaluation could occur for any given local decision to see whether in that instance there is entailment between the black-box model and the xAI model. As we saw, one general downside of MAP is that an input-output pairing does not capture which features xAI methods determine are (ir)relevant ELIMINATION captures this aspect of explainability. Since ELIMINATION can be evaluated per local decision, there is more flexibility for success. Some local decisions may not satisfy ELIMINATION, while others do. Indeed, in the cases where the xAI method works well, ELIMINATION should be satisfied. But the trade-off here is that idealization failure would needs to be tested for each local decision.

Lastly, as yet another alternative, some philosophers of science have argued only a probabilistic notion of (causal) relevance or difference-making is necessary for MinI [77]. Prob can apply to any other rule. It says that the probability that a rule applies to x is greater than some probability threshold *t*. That said, PROB may not align well with the ideals for MinI for many xAI purposes, for example users may not want to know what the most probable reason for the decision was, but the actual reason for the decision. However, when using xAI for the purposes of de-bugging or de-biasing a black-box model, PROB could be appropriate.

4.3.2 *Role and Limits of Current Work.*

The IDEALS and RULES phase is the area in xAI idealization evaluation that requires innovation. In philosophy, Fleisher [29] argues feature importance methods are a kind of MinI, but he argues this on the level of IDEALS and stops short of discussing whether particular xAI methods actually succeed at MinI instead of merely *aiming* for MinI. Other works focus on the norms and ideals of *explanations* for xAI. Citing one example, Watson et al. [109] propose the ideal of sufficiency for xAI methods because they provide potentially more useful and 'lower cost' explanations for users. However, the norms and ideals for explanation are distinct from the norms and ideals that govern idealization evaluation. Idealization evaluation, first and foremost, treats xAI models not as explanation tools, but as idealized models of more complex models. SIDEs evaluates whether the distortions a given xAI method engages in succeeds at living up to the purpose and norms of idealization theory.

On the level of RULES, there are existing experimental techniques that have been used to evaluate LIME and SHAP that capture some of the spirit of our suggested entailment rules for MinI, like MAP discussed above. While LIME and SHAP do not satisfy MAP because the accuracy rates do not reach perfect alignment between the black-box model and the xAI model, MAP's PROB counterpart has some level of success depending on the probability threshold. However, even a .8 probability may be too low to establish a strong sense of causal entailment for MinI. So if the ideal for xAI is a causal ideal then there is still idealization failure in the leading feature importance methods. There is no experimental test that captures ELIMINATION that I am aware of. However, feature importance methods are vulnerable to adversarial manipulation. Slack et al. [92] were able to create explanations that hid the most salient feature for classification for SHAP and LIME. Ghorbani et al. [32] found such methods were highly sensitive to small changes input data. Others have found that they are not able to capture causal notions [59, 76, 85]. This type of manipulation suggests that feature importance methods are not robustly conforming to ELIMINATION. Moreover, adversarial examples where the adversarial classifier achieves strong PROB(MAP) but the perturbed instances are different (see [92]) shows the tension between more global oriented rules like MAP and local entailment rules like ELIMINATION. This suggests that in order to fully capture MinI, it may be necessary to satisfy both MAP and ELIMINATION.

Innovation on new experimental evaluations that instantiate idealization rules could be promising. In this paper, I took just one conception of (causal) entailment from Strevens to ground potentially new experimental evaluation tests for xAI. XAI researchers should consider the idealization norms and ideals they are aiming to achieve and align experimental evaluation tests with these norms. Existing theories of idealization in philosophy of science could serve as inspiration for establishing quantitative evaluation metrics for idealization.

Key questions for the IDEALS and RULE phase are:

- *What are the norms and values that govern and justify an idealization practice?*
- *What are the possible ways to operationalize ideals to experimentally and formally evaluate whether an idealized model satisfies the ideals of the idealization practice?*
- *What are the trade offs between different rules?*
- ***Example**: Different possible entailment rules for MinI and the limitations of each.*
- ***Success**: RULES adequately reflect IDEALS. The xAI method satisfies the rule by passing an experimental test.*

4.4 User-facing explanations

Work on idealization in the natural sciences considers scientists as stakeholders. The ideal gas law idealization is successful mainly because the intended audience generally knows enough about physics to understand where the idealizations lie. XAI, on the other hand, serves many diverse stakeholders, many of which do not know how ML works in any detail. Therefore, attention must be paid to the way idealizations are presented to different stakeholders through USER-FACING explanations. Thus, the last step for the SIDEs framework is evaluating user-facing explanations.

4.4.1 *Evaluating USER-FACING explanations.*

Evaluating the explanations that target users receive involves user testing and user studies to ensure that explanations align with PURPOSE and do not mislead users with its idealizations. One general pitfall for aligning user-facing explanations with user values is the potential for 'explanation hacking' where xAI methods are so flexible to display only those that are agreeable to users [99], which can result in fairwashing [1]. Even if the xAI method is a successful idealization from a scientific or mathematical perspective (i.e. passes the RULE phase), there can be idealization failure if it misleads users through explanation hacking or fairwashing techniques. Moreover, Mittelstadt et al. [65] found that users found LIME and SHAP unintuitive. Other considerations include doing user-study research not just for experts, but for a variety of users that have different background assumptions [26]. Including users from the global south, which are often ignored [71, 75].

4.4.2 *Role and Limits of Current Work.*

Several existing works address issues regarding user-facing xAI. Some provide frameworks that incorporate stakeholder interests [22, 44, 54, 66, 114], others explore how different explanation types affect user-trust [41], cognitive bias [11, 18], and understanding [19, 68]. SIDEs adds to this by highlighting the need for user-facing explanations to align with the purpose of *idealizations*, conveying the ideals and norms of the idealizations in the explanations they receive. For example Schneider and Vlachos [87], used language similar to MinI when describing the results of their user study, saying that users could tease out what was relevant and irrelevant to a model decision.

Key questions for the USER-FACING EXPLANATION phase are:

- *Do users understand the purpose of the explanation?*
- *How can user-facing explanations convey that the explanation is an idealization?*
- *Are user-facing explanations evaluated in terms of the purpose? Are they evaluated in terms of another purpose?*
- ***Example**: A user study asks users about their impressions regarding the purpose and limits to the explanation.*
- ***Success**: User-facing explanations align with PURPOSE and users acknowledge the limited scope of explanations.*

5 EVALUATING COUNTERFACTUAL EXPLANATION METHODS

In the previous section, I introduced the SIDEs framework and presented a qualitative evaluation of leading feature importance methods, LIME and SHAP. In this section, I identify risks of idealization failure in counterfactual explanation (CE) methods, using SIDEs. It is important to note that proponents of CE methods boast that CE cannot be false since counterfactuals are generated from the black-box model itself [68]. However, such methods still engage in idealization. CE methods must select which counterfactual scenarios are the most salient from a larger set of possible counterfactuals and implicitly rely on notions of difference-making that seek to tease out the relevant counterfactual scenarios from the less relevant. This is one reason why it is important to re-conceptualize xAI as seeking to solve an idealization problem, instead of the current frame of xAI aiming at 'faithful' explanations.

***Misalignment in CE:**.* Wachter et al. [107] highlight three different purposes for counterfactual explanation methods: *i)* explain why a decision was reached, *ii)* provide grounds to contest the decision, and *iii)* provide users with actionable changes to reverse the decision. SIDEs first requires that the idealizations used align with each specific purpose. *i)* has a clear epistemic purpose aimed at gaining understanding of the black-box model behavior, while using CE for *iii)*–known as recourse–has gained considerable attention regarding its ethical promise [101, 105]. A recourse explanation is one where users are given feasible actions for them to undertake to reverse a model decision (e.g. paying down existing debt to qualify for a loan). Importantly, recourse and purely epistemic explanatory aims come apart [47, 99]. Since recourse provides users with actionable changes that they can make to reverse a decision, recourse explanations can mask bias and principle-reason explanations [5]. It could be that an immutable feature, like gender or race, was *the* biggest difference-maker for why the model made its decision. Such an explanation cannot, in principle, be a recourse explanation. This is one reason why Sullivan and Verreault-Julien [101] suggest conceptualizing recourse as a recommendation to avoid misleading users.

The type of idealizations that can satisfy MinI for *understanding the decision* do not immediately translate to idealizations that are acceptable if the purpose of the idealization is to provide users with recourse. For example, for the epistemic purpose of understanding, CE methods can idealize and distort the underlying causal structure in the data and idealize away any interdependence between features–especially if the underlying black-box model ignores interdependence between features. In the epistemic case, satisfying MinI only requires alignment between the xAI model and the black-box model (i.e. answering the MODEL-MODEL question). However, recourse explanations have a different target: the relationship between model features *and the world* [42], thereby aiming to answer a MODEL-WORLD question. For recourse, an idealization that ignores feature interdependence and the underlying causal structure in data will likely fail MinI RULES that are calibrated to capture aspects of real-world causal efficacy. Indeed, works have criticised CE methods that ignore feature interdependence as creating unrealistic advice [39].

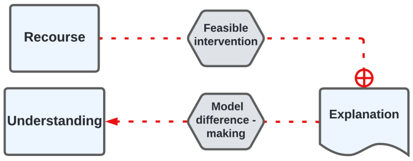

Figure 2: Recourse Alignment Failure

PURPOSE-alignment-failure can carry over to the way xAI explanations are conveyed to users in the USER-FACING EXPLANATION phase. For example, recourse explanations are often presented as answering a MODEL-MODEL question (i.e. how the black-box model behaves), with several works evaluating their recourse method on whether users have similar understanding of the models decision boundary compared to, e.g. LIME [68, 103, 107]. However, since the underlying purposes of recourse CE are feasibility and actionability, evaluating recourse CE should be done in terms of whether users find the recourse CE feasible. Understanding the model's decision is secondary. Figure 2 shows this type of recourse CE alignment failure, where the explanation is evaluated based on the wrong PURPOSE. This is not to say that recourse explanations could not be successful idealizations. SIDEs maintains that omitting or distorting the central reasons for a model's decision from a recourse explanation is legitimate so long as the USER-FACING EXPLANATION is aligned with actionability, while making clear to the user it is not an *epistemic* explanation. Of course, it is possible that CE could satisfy both an epistemic purpose and recourse. However, for an idealization to achieve both aims there is a considerably higher bar where the CE method would need to satisfy both difference-making w.r.t. the MODEL-MODEL question and difference-making w.r.t. the MODEL-WORLD question. We should expect that this might be possible in some cases, but not likely in cases where immutable features are the largest difference-maker for a model's decision.

IDEAL and RULE failure. Many of the same issues that come up with RULE evaluation for feature importance methods also appear with CE methods. However, like the alignment issues discussed above, recourse CE has unique risks. The governing IDEALS for recourse are feasibility and actionability. Thus, a structural causal model (SCM) that captures how features within a model causally dependent on each other [6, 50, 74] will be necessary to satisfy MinI, albeit an idealized SCM. However, SCMs are far from attainable, requiring a link to the causal realities outside of the model [6, 96]. Thus, the PURPOSE of recourse, coupled with the IDEALIZATION PRACTICE MinI, requires a causal RULE that can uphold the IDEAL of a SCM. However, Karimi et al. [42] find most works on recourse do not even aim for a SCM. Thus, these methods are engaging in idealization failure. Karimi et al. [43], on the other hand, employ a probabilistic approach to try and capture the ideal of a SCM with imperfect causal knowledge, and thus is a candidate for a PROB(SCM) rule, and might indeed be an idealization success, depending on how the PROB(SCM) rule isolates difference-makers.

All told, researchers need to be mindful that MODEL-WORLD questions and MODEL-MODEL questions require different idealizations and have different idealization standards.

6 TURNING IDEALIZATION FAILURE INTO SUCCESS

In our limited evaluation of feature importance and CE methods we found that on the working hypothesis that xAI methods are aiming for MinI there is widespread idealization failure, due to misalignment with PURPOSE and RULE failure, suggesting that these methods may likely distort more than just the non-difference-makers. Where does this leave us? In this section I discuss possible remedies for idealization failure and areas for future research.

Adopting a different idealization practice: If it is too difficult to satisfy one idealization practice, such as MinI, then idealization success can occur by creating alignment with a different idealization practice. RULE failure under one idealization practice does not entail RULE failure under another. Figure 3 shows realignment with a different idealization practice after RULE failure. As we said at the outset there are a number of idealization practices that philosophers of science have discussed. One alternative idealization practice that may be well suited for xAI is multiple-model idealization (MMI) [110]. For some phenomena there may be several trade-offs that makes it either impractical or epistemically lacking to rely on just one model. MMI involves multiple models, with each model capturing one aspect or trade-off for some phenomenon. For example, chemists use both valence bond and molecular orbital models despite their incompatible assumptions [110]. The underlying justification for MMI relies on the impossibility (either practically or necessarily) of a single model capturing maximal goals of representation, such as accuracy and generality [57]. Furthermore, through robustness techniques among multiple models, MMI teases out relevance from irrelevance, something that MinI might fail to achieve in a single model [111]. While each individual model in the MMI might have drawbacks, the collection of models taken together is able to provide understanding.

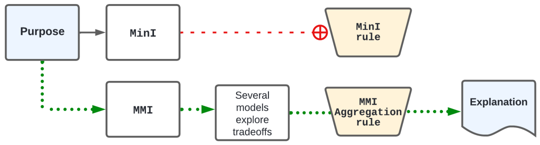

Figure 3: Adopting an alternative idealization practice to address rule failure

One notable solution to the idealization failure with MinI is to use LIME, SHAP and CE as part of a MMI instead. Realigning to MMI means that the rules for each xAI method can be less demanding. Each model, fulfiling a different ideal or rule explores different trade-offs, together capturing how a black-box model makes decisions. Moreover, MMI can be a temporary solution to xAI manipulation. Slack et al. [92] found that adversarial attacks designed for LIME were ineffective against SHAP. Even though they also found that adversarial attacks designed against SHAP affected LIME, if the MMI also includes CE methods (or other methods) the vulnerability that each will be affected by the same attack diminishes. Furthermore, MMI can help with the problem of multi-purposes for xAI models. Users can be provided multiple explanations to fulfil these multiple purposes. However, MMI is not an simple fix. First, it is important to develop an aggregation rule for MMI idealizations. How should we weigh the different and sometimes conflicting models? This is a considerable undertaking. Second, current work has discussed that providing users with multiple different explanations can be counter-productive, creating confusion and cognitive overload [79]. Thus, MMI might not be a useful idealization practice outside of an engineering model-auditing setting.

xAI, a novel idealization practice? In this paper, we looked at mature theories of idealization from the philosophy of science that were developed with the natural sciences in mind. It would not be surprising if those theories are altogether ill-suited for xAI since xAI and ML research has very different and specific requirements compared to the natural sciences. For example, one unique aspect of many xAI methods are their *hyper locality*. In the natural sciences, idealizations often move away from local particulars to more global generalities. But current methods of xAI are distorting the global generalities of the black-box model to zoom in on a particular local explanation. Perhaps this is a novel *hasty generalization* idealization practice. One area for future research is developing what this potentially novel idealization practice consists of and how it might be justified and grounded as a legitimate idealization practice. Importantly, the SIDEs framework cautions against idealization success simply by fiat (i.e. by claiming a new idealization practice). SIDEs requires a justification step for motivating why such a practice is legitimate.

7 CONCLUSION

XAI methods have received their fair share of criticism. However, with this paper I argued that one type of criticism–that xAI methods produce false explanations of black-box models–deserves closer attention. Specifically, this paper seeks to animate a new interdisciplinary research program in xAI that develops a theory of xAI idealizations and idealization evaluation. It is not simply departure from the truth that is problematic, but idealization failure. I introduced the SIDEs framework as a way for researchers to separate successful idealizations from deceptive explanations. SIDEs is a generalizable and modular conceptual framework that can guide researchers with key questions for reflection and qualitative evaluation, as well as provide the normative foundation for developing more concrete evaluative tests and benchmarks for xAI methods. SIDEs is primarily aimed at xAI researchers when developing and evaluating their methods (esp. the ideals and rules phase). However, there is also a place for SIDEs in more downstream uses. Those who deploy xAI models could use SIDEs in selecting which xAI method would be more successful for their purpose, such as providing recourse explanations vs. model auditing. However, this would require clear guidance on the results of the rest of the SIDEs workflow from xAI researchers. Moreover, the user-facing explanation phase could be useful for thinking about how to comply with the right to explanation in AI governance.

Using SIDEs, we found considerable risks of idealization with leading xAI methods. There are many ways in which current work in xAI is useful for idealization evaluation, such as identifying the purpose of xAI methods, developing user-centered studies to evaluate the efficacy of explanations. However, there are central ways in which innovation is necessary: 1) identifying and solidifying an idealization practice for xAI, including a justification for that practice, and 2) developing experimental tests that aim at evaluating how well an xAI method idealizes the target black-box model.

ACKNOWLEDGMENTS

For helpful comments and discussions I'd like to thank Céline Budding, Thomas Grote, Yeji Streppel, Philippe Verreault-Julien, Carlos Zednik, with special thanks to John Mumm. I presented this work in various stages in development at the Munich Center for Mathematical Philosophy, University of Tübingen, Utrecht University, and Nicolaus Copernicus University. I thank those in attendance for the fruitful discussions that followed. This work is supported by the Netherlands Organization for Scientific Research (NWO grant number VI.Veni.201F.051).

REFERENCES

[1] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.
[2] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).
[3] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104* (2017).
[4] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1194–1206.

[5] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.

[6] Sam Baron. 2023. Explainable AI and causal understanding: Counterfactual approaches considered. *Minds and Machines* 33, 2 (2023), 347–377.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[8] Robert W Batterman. 2001. *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence.* Oxford University Press.

[9] Robert W Batterman and Collin C Rice. 2014. Minimal model explanations. *Philosophy of Science* 81, 3 (2014), 349–376.

[10] Claus Beisbart and Tim Räz. 2022. Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass* 17, 6 (2022), e12830.

[11] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*. 78–91.

[12] Florian J Boge. 2022. Two dimensions of opacity and the deep learning predicament. *Minds and Machines* 32, 1 (2022), 43–75.

[13] Alisa Bokulich and Wendy Parker. 2021. Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science* 11 (2021), 1–26.

[14] Oliver Buchholz. 2023. A Means-End Account of Explainable Artificial Intelligence. *Synthese* 202, 2 (2023), 33.

[15] Stefan Buijsman. 2022. Defining explanation and explanatory depth in XAI. *Minds and Machines* 32, 3 (2022), 563–584.

[16] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.

[17] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[18] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions under Different Levels of Uncertainty. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–42.

[19] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[20] Kathleen A Creel. 2020. Transparency in complex computational systems. *Philosophy of Science* 87, 4 (2020), 568–589.

[21] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems* 32 (2019).

[22] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[23] Eamon Duede. 2022. Deep learning opacity in scientific discovery. *Philosophy of Science* (2022), 1–13.

[24] Juan M Durán and Nico Formanek. 2018. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines* 28 (2018), 645–666.

[25] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (2021), 329–335.

[26] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[27] Catherine Z Elgin. 2017. *True enough.* MIT press.

[28] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. 2021. What is interpretability? *Philosophy & Technology* 34, 4 (2021), 833–862.

[29] Will Fleisher. 2022. Understanding, idealization, and explainable AI. *Episteme* 19, 4 (2022), 534–560.

[30] Timo Freiesleben and Gunnar König. 2023. Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In *World Conference on Explainable Artificial Intelligence*. Springer, 48–65.

[31] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[32] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3681–3688.

[33] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[34] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[35] Thomas Grote. 2021. Trustworthy medical AI systems need to know when they don't know. *Journal of medical ethics* 47, 5 (2021), 337–338.

[36] Thomas Grote, Konstantin Genin, and Emily Sullivan. 2024. Reliability in Machine Learning. *Philosophy Compass* (2024).

[37] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating Counterfactual Explanations with Natural Language. In *ICML Workshop on Human Interpretability in Machine Learning*. 95–98.

[38] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, and Simine Vazire. 2021. Integrating explanation and prediction in computational social science. *Nature* 595, 7866 (2021), 181–188. https://doi.org/integrative-modeling-2021

[39] Giles Hooker, Lucas Mentch, and Siyu Zhou. 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing* 31 (2021), 1–16.

[40] Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 279–287.

[41] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[42] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).

[43] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems* 33 (2020), 265–277.

[44] Atoosa Kasirzadeh. 2021. Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 14–14.

[45] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 228–236.

[46] Benedikt Knüsel and Christoph Baumberger. 2020. Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A* 84 (2020), 46–56.

[47] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. 2023. Improvement-focused causal recourse (ICR). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11847–11855.

[48] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* (2022).

[49] Maya Krishnan. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* 33, 3 (2020), 487–502.

[50] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[51] Himabindu Lakkaraju and Osbert Bastani. 2020. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[52] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[53] Marc Lange. 2016. *Because Without Cause: Non-Casual Explanations In Science and Mathematics.* Oxford University Press.

[54] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[55] Insa Lawler. 2021. Scientific understanding and felicitous legitimate falsehoods. *Synthese* 198, 7 (2021), 6859–6887.

[56] Breiman Leo. 2001. Statistical modeling: The two cultures. *Statistical science* 16, 3 (2001), 199–231.

[57] Richard Levins. 1966. The strategy of model building in population biology. *American scientist* 54, 4 (1966), 421–431.

[58] David Kellogg Lewis. 1973. Counterfactuals. (1973).

[59] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[60] Alex John London. 2019. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report* 49, 1 (2019), 15–21.

[61] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[62] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable Credit Application Predictions With Counterfactual Explanations. In *NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*.

[63] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[64] Joseph Millum and Danielle Bromwich. 2021. Informed consent: What must be disclosed and what must be understood? *The American Journal of Bioethics* 21, 5 (2021), 46–58.

[65] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[66] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[67] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[68] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[69] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164* (2022).

[70] Rune Nyrup and Diana Robinson. 2022. Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics and information technology* 24, 1 (2022), 13.

[71] Chinasa T Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI explainable in the Global South: A systematic review. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*. 439–452.

[72] Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines* 29, 3 (2019), 441–459.

[73] Wendy S Parker. 2020. Model evaluation: An adequacy-for-purpose view. *Philosophy of Science* 87, 3 (2020), 457–477.

[74] Judea Pearl. 2009. *Causality*. Cambridge university press.

[75] Uwe Peters and Mary Carman. 2024. Cultural Bias in Explainable AI Research: A Systematic Analysis. *Journal of Artificial Intelligence Research* 79 (2024), 971–1000.

[76] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. 2018. Model agnostic supervised local explanations. *Advances in neural information processing systems* 31 (2018).

[77] Angela Potochnik. 2015. Causal patterns and adequate explanations. *Philosophical Studies* 172, 5 (2015), 1163–1182.

[78] Angela Potochnik. 2017. *Idealization and the Aims of Science*. University of Chicago Press.

[79] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[80] Corinne Purtill and Quartz. 2015. Behold, the Geographically Accurate Tube Map. *The Atlantic* (2015). https://www.theatlantic.com/entertainment/archive/2015/09/behold-the-geographically-accurate-tube-map/405967/

[81] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[82] Collin Rice. 2019. Models don't decompose that way: A holistic view of idealized models. *The British Journal for the Philosophy of Science* 70, 1 (2019), 179–208.

[83] Yasha Rohwer and Collin Rice. 2013. Hypothetical pattern idealization and explanatory models. *Philosophy of Science* 80, 3 (2013), 334–355.

[84] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2022. Towards Human-centered Explainable AI: User Studies for Model Explanations. *arXiv preprint arXiv:2210.11584* (2022).

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

[86] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.

[87] Johannes Schneider and Michalis Vlachos. 2023. Explaining classifiers by constructing familiar concepts. *Machine Learning* 112, 11 (2023), 4167–4200.

[88] Andrew Selbst and Julia Powles. 2018. "Meaningful information" and the right to explanation. In *conference on fairness, accountability and transparency*. PMLR, 48–48.

[89] Elay Shech. 2018. Infinite idealizations in physics. *Philosophy Compass* 13, 9 (2018), e12514.

[90] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

[91] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems* 34 (2021).

[92] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[93] Michael Strevens. 2011. *Depth: An account of scientific explanation*. Harvard University Press.

[94] Michael Strevens. 2016. How idealizations provide understanding. In *Explaining understanding*. Routledge, 53–65.

[95] Michael Strevens. 2019. The structure of asymptotic idealization. *Synthese* 196, 5 (2019), 1713–1731.

[96] Emily Sullivan. 2022. Understanding from machine learning models. *The British Journal for the Philosophy of Science* 73, 1 (2022), 109–133.

[97] Emily Sullivan. 2023. Do Machine Learning Models Represent Their Targets? *Philosophy of Science* (2023), 1–11.

[98] Emily Sullivan, Dimitrios Bountouridis, Jaron Harambam, Shabnam Najafian, Felicia Loecherbach, Mykola Makhortykh, Domokos Kelen, Daricia Wilkinson, David Graus, and Nava Tintarev. 2019. Reading news with a purpose: Explaining user profiles for self-actualization. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 241–245.

[99] Emily Sullivan and Atoosa Kasirzadeh. 2024. Explanation Hacking: The perils of algorithmic recourse. In *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library.

[100] Emily Sullivan and Kareem Khalifa. 2019. Idealizations and understanding: Much ado about nothing? *Australasian Journal of Philosophy* (2019).

[101] Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 712–722.

[102] John Symons and Ramón Alvarado. 2022. Epistemic injustice and data science technologies. *Synthese* 200, 2 (2022), 87.

[103] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.

[104] Ronald Veldhuizen. 2022. Het stikstofmodel van het RIVM bevindt zich in het centrum van een crisis: hoe werkt het eigenlijk? *de Volkskrant* (2022). https://www.volkskrant.nl/wetenschap/het-stikstofmodel-van-het-rivm-bevindt-zich-in-het-centrum-van-een-crisis-hoe-werkt-het-eigenlijk~b31a0dc8/

[105] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[106] Kate Vredenburgh. 2022. The right to explanation. *Journal of Political Philosophy* 30, 2 (2022), 209–229.

[107] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018).

[108] David S Watson. 2022. Conceptual challenges for interpretable machine learning. *Synthese* 200, 1 (2022), 1–33.

[109] David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. 2021. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Uncertainty in Artificial Intelligence*. PMLR, 1382–1392.

[110] Michael Weisberg. 2007. Three kinds of idealization. *The journal of Philosophy* 104, 12 (2007), 639–659.

[111] William C. Wimsatt. 2012. Robustness, Reliability, and Overdetermination (1981). In *Characterizing the Robustness of Science*. 61–78.

[112] Markus Wolfensberger and Anthony Wrigley. 2019. *Trust in Medicine*. Cambridge University Press.

[113] James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

[114] Carlos Zednik. 2021. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology* 34, 2 (2021), 265–288.