Synthese (2023) 202:109
https://doi.org/10.1007/s11229-023-04334-9

ORIGINAL RESEARCH

Beyond generalization: a theory of robustness in machine learning

Timo Freiesleben[1] · **Thomas Grote**[1]

Received: 17 January 2023 / Accepted: 5 September 2023 / Published online: 27 September 2023

Abstract

The term *robustness* is ubiquitous in modern Machine Learning (ML). However, its meaning varies depending on context and community. Researchers either focus on narrow technical definitions, such as adversarial robustness, natural distribution shifts, and performativity, or they simply leave open what exactly they mean by robustness. In this paper, we provide a conceptual analysis of the term *robustness*, with the aim to develop a common language, that allows us to weave together different strands of robustness research. We define robustness as the relative stability of a robustness target with respect to specific interventions on a modifier. Our account captures the various sub-types of robustness that are discussed in the research literature, including robustness to distribution shifts, prediction robustness, or the robustness of algorithmic explanations. Finally, we delineate robustness from adjacent key concepts in ML, such as extrapolation, generalization, and uncertainty, and establish it as an independent epistemic concept.

Keywords Robustness · Machine Learning · Generalization · Models in Science · Uncertainty · Extrapolation

Timo Freiesleben and Thomas Grote have contributed equally to this work.

✉ Timo Freiesleben
timo.freiesleben@uni-tuebingen.de

Thomas Grote
thomas.grote@uni-tuebingen.de

[1] Cluster of Excellence: "Machine Learning: New Perspectives for Science", University of Tübingen, Maria-von-Linden-Straße 6, 72076 Tübingen, Germany

1 Introduction

Machine Learning (ML) models face a common problem: they achieve astonishing predictive performance under testing conditions; yet, they often fail in real-world deployment. To motivate the problem, consider some examples[1]:

- **Shortcuts**: An ML model used to diagnose pneumonia from X-rays has been found to achieve its predictive performance by exploiting confounding information—e.g., hooks at the edges of the images. When transferred to real-world clinical settings, this is likely to result in diagnostic errors (Zech et al., 2018).
- **Natural Distribution Shifts**: Despite performing as well as certified ophthalmologists in the diagnosis of diabetic retinopathy during training, a clinical field study in Thailand showed that the ML model could not grade many of the fundus images, since they were made under non-ideal lighting conditions (Beede et al., 2020).
- **Adversarial Attacks**: Small targeted manipulations—such as placing a colored patch in the scene—have paralyzed the visual flow system of self-driving cars (Ranjan et al., 2019).
- **Performativity**: Although the predictive performance of an ML model for estimating loan applicants' default risks was initially fair across different demographics, after a few months of use, the model disadvantages marginalized social groups by assigning them higher interest rates. This is because more privileged social groups have learned to game its functionality (Perdomo et al., 2020).

In each of these cases, the remedy is to improve the ML model's robustness. But what does it mean for an ML model to be *robust*? Despite being a key desideratum in both, ML research and ML deployment, the notion itself is not well-defined. As a case in point, it does not appear in any of the indexes of the highly influential textbooks by Hastie et al. (2009), Shalev-Shwartz and Ben-David (2014), or Goodfellow et al. (2016). Survey and benchmark papers on ML robustness rather focus only on specific robustness notions, such as adversarial robustness (Dong et al., 2020; Dreossi et al., 2019; Serban et al., 2020), robustness to natural distribution shifts (Drenkow et al., 2021; Hendrycks et al., 2021; Koh et al., 2021; Taori et al., 2020), or shortcut learning (Geirhos et al., 2020). Evidently, robustness research in ML is conceptually fragmented.

Given that the concept of robustness is well-researched in philosophy of science (Levins, 1966; Orzack & Sober, 1993; Schupbach, 2018; Weisberg, 2006; Wimsatt, 1981; Woodward, 2006), one might expect to find a ready-made general theory that can be straightforwardly applied to robustness in ML. However, there are two obstacles: Firstly, the debate on robustness analysis (RA) in science often focuses on parsimonious analytical models that do not involve a learning process, which is why RA is unable to capture the specifics of complex ML models. Secondly, although there is considerable methodological overlap between RA in computer simulations (CS) for scientific inquiry (Boge, 2021; Durán & Formanek, 2018; Gueguen, 2020; Parker, 2011) and robustness research in ML, the former strand tends to tailor its theoretical frameworks to specific cases from High-Energy Physics or climate science. This

[1] The notions refer to robustness sub-types that we explain in Sect. 3.3.

domain-specificity makes it challenging to apply theories of robustness from CS to ML.

In this paper, we provide a unifying theory of robustness for ML and showcase how it facilitates explaining and synthesizing different robustness sub-types in ML. We define robustness as the relationship between two entities, a robustness target and a robustness modifier, where the target remains (relatively) stable under relevant interventions on the modifier (see Sect. 3). We identify crucial robustness targets and modifiers within the ML pipeline and discuss strategies for ensuring robustness with respect to these targets/modifiers. Our focus here is on the robustness of the model's deployment performance under distribution shifts. Finally, we highlight what distinguishes robustness from adjacent concepts in ML, such as extrapolation, generalization, and uncertainty quantification (see Sect. 4). In particular, we argue that robustness is an independent epistemic notion, presupposing the generalization capacities of the model under training conditions, but going beyond that by shifting the focus to the reliable *real-world* deployment of ML models.

Since ML models have become mainstays in many scientific fields, such as computational neuroscience (Kriegeskorte & Douglas, 2018) or climate research (Ham et al., 2019), we hope that our work also helps clarifying the requirements for scientific inference with ML models, in addition to providing guidance for debates on ML regulation.

To set the stage, Sect. 2 discusses the conceptual fragmentation of robustness in ML, while also pointing out discontinuities between *traditional* RA in the sciences and ML. Section 3 develops an account of robustness and shows how this account captures different robustness sub-types across the ML pipeline. Finally, Sect. 4 delineates robustness from adjacent concepts in ML and statistics and addresses the normative implications, resulting from the revaluation of robustness.

2 The (dis)unity of robustness in machine learning

In this section, we discuss the conceptual fragmentation of robustness research in ML and situate our work within the philosophical robustness literature.

2.1 The conceptual fragmentation of robustness

Although robustness is ubiquitous in ML research and societal debates on ML regulation, the concept itself is under-analyzed. This is surprising, to say the least. Like in the debate on ML *interpretability* (Creel, 2020; Lipton, 2018; Sullivan, 2022; Watson, 2022), where the central concept lacks a clear definition, we believe that it is most useful to start by tackling the conceptual foundations of robustness. Clarifying the notion(s) of robustness is a prerequisite for ML researchers to get a better grip on (i) what constitutes adequate evaluation criteria, (ii) where novel techniques to improve robustness are needed, and (iii) what guarantees can be meaningfully provided for robustness of ML models.

What explains the conceptual fragmentation of robustness in ML? As we see it, there are three possible conceptual explanations:

- **Variety of Purposes:** There may be different reasons and deployment scenarios why and where robustness is desirable. In turn, this leads to a focus on different subtypes of robustness.[2] For example, warding off adversarial attacks can be deemed pivotal for self-driving cars but irrelevant to an epidemiologist, estimating virus spread with ML models. In contrast, for the epidemiologist, remedying effects arising from model performativity (see Sect. 3.3) is crucial.
- **Categorical Heterogeneity:** Robustness could be a vague umbrella term, denoting a bundle of distinct phenomena without an underlying shared structure. Consequently, a unified theory of robustness is destined to miss its mark.
- **Reducibility:** Perhaps robustness can be replaced by or reduced to other more refined concepts from ML or statistics, such as generalization, extrapolation, or uncertainty. If this is the case, little is gained by analyzing robustness as a separate concept.

Our goal of establishing robustness as an independent epistemic concept and unifying different strands of robustness research in one framework is compatible with the *variety of purposes* explanation: throughout the paper, we showcase how different strands of ML research, concerned with robustness in ML, are motivated by different desiderata. That being said, we reject the latter two explanations: we will undermine the *categorical heterogeneity* explanation in Sect. 3 by providing an account of robustness that captures a wide variety of existing robustness sub-types. Moreover, we will address the *reducibility*-explanation in Sect. 4 by arguing that robustness cannot be reduced to the most promising candidate concepts, while also highlighting its specific epistemic functions.

The list of possible explanations is not exhaustive. For example, there might be also sociological factors, which we will not pursue in this paper, such as the *hyper-specialization* of ML research: ML research is progressing rapidly, with theory lagging far behind practical breakthroughs. To publish at major ML conferences, it is strategic to focus on concise application problems rather than, say, an overarching theory of robustness. Viewed in this light, conceptual fragmentation would be a side effect of research and publication dynamics.

2.2 Robustness analysis in the sciences

The most refined philosophical discussions of the concept of robustness can be found in the debate on RA in the sciences. RA describes a family of techniques that allow discerning whether some phenomenon of interest is invariant, using independent (or changing) means of detection (Schupbach, 2018; Wimsatt, 1981). This acts as a safeguard against competing explanations of a result's occurrence.

[2] A comparison case is again the interpretable ML (IML) literature (Molnar, 2020), where it is well-established to distinguish between different purposes (e.g., recourse or scientific inference) and explanation-targets (e.g., the model itself or the modeled phenomenon), resulting in different epistemic requirements (Creel, 2020; Freiesleben et al., 2022; Zednik, 2021).

The relevant account of RA is fairly liberal: it can be applied to a number of different phenomena of interest, such as theorems, hypotheses, experimental setups, model performance, and so on (Schupbach, 2018; Wimsatt, 1981). RA is particularly valuable in disciplines that lack a unifying theoretical framework, such as in the life sciences, or in disciplines that rely on highly idealized modeling assumptions, like in economics, where it is difficult to empirically test whether the target system is sufficiently well represented (Kuorikoski et al., 2010). Evidently, it is also highly relevant for ML.

To understand the specifics of RA for ML, we start by considering RA for analytical models as a contrast class. Here, the basic methodology (often called 'triangulation') goes as follows (Weisberg, 2006): First, identify a group of sufficiently diverse models that all predict a common result.[3] Second, investigate whether all these models share a robust property, such as the reliance on a specific variable or a generalizing theoretical property that can be proven. Third, provide interpretations of how the robust mathematical strands in a model capture the causal structure of a given empirical phenomenon of interest. Finally, formulate a robust theorem, which has the form:

> Ceteris paribus, if [common core (causal) structure] obtains, then [robust property] will obtain.
> (Weisberg, 2006, p. 731)

As an illustration, the fact that greenhouse gas causation is an integral part of many climate prediction models underscores that greenhouse gas concentration is causally relevant for real-world increases in global warming (Lloyd, 2010, p. 982).

Against this backdrop, it is important to highlight some glaring differences between the methodology of RA for analytical models and ML. Firstly, ML is an instrumentalist approach that prioritizes predictive performance over representational accuracy.[4] Whether a model's decision-function captures the structure of the target system matters only insofar as it enables greater predictive capacities.[5] Secondly, the opacity of ML models (Boge, 2022; Creel, 2020; Sullivan, 2022) makes it challenging to achieve a degree of transparency, necessary for detecting common model properties. Finally, rather than comparing a set of sufficiently independent models to extract robust model properties, ML aims to establish the robustness of an *individual* model by testing its performance across sufficiently diverse data scenarios. Even when there is such a set of diverse predictive models in ML, they are typically merged into an ensemble—to achieve even greater predictive accuracy and improved robustness—rather than scrutinizing their commonalities, (Zhang & Ma, 2012).

[3] Note, however, that Schupbach (2018) criticizes appeals to evidential diversity in RA on the grounds that all prevailing formal explications of such diversity are unsatisfactory.

[4] Indeed, one of the key philosophical questions concerning ML in the sciences is whether deep neural networks in particular possess meaningful representational properties. Strictly speaking, there are two issues that need to be disentangled: (i) whether ML models possess representational capacities that are akin to mental representations; (ii) whether ML models provide meaningful scientific representations of the target system. Concerning (i), a good starting point, with a positive outlook, is Buckner (2018), whereas Boge (2022) is a good starting point for (ii), albeit with a pessimistic view. We thank an anonymous reviewer for this helpful comment.

[5] Although an alignment of the model's structural properties with the target system can be crucial for enabling interpretability (Sullivan, 2022).

Things get thornier when we compare RA for ML and CS. Due to the advent of CS in fields such as High Energy Physics or climate science, the challenge is to determine when we can trust a simulation model to generate credible experimental results and under what conditions we are permitted to infer that a CS model faithfully represents its target system (Boge, 2021; Durán & Formanek, 2018; Karaca, 2022).

Just like in ML, RA for CS also strongly centers around predictive performance. Indeed, despite differences in their model architectures, highly parameterized CS and ML models might both be deemed to be instrumental devices that are used to make accurate predictions (Boge, 2022). It is therefore not sufficient to consider (internal) model properties to establish robustness of simulation models.

Consequently, a common means to determine the robustness of CS in the context of climate science is to validate the model against historical climate events (Heiri et al., 2014). Relatedly, in the Intergovernmental Panel on Climate Change (IPCC) Assessment Report (currently in its 6th iteration), a key component of the RA is to examine the performance spread of a model ensemble—for the purpose of estimating the uncertainty of climate simulations (Arias et al., 2021, pp. 48–51).

Karaca (2022) emphasizes the robustness of experimental procedures to data selection, relating to the robustness of a specific means of detection with respect to the inputs we feed to them. This seems conceptually closely connected to forms of adversarial robustness, where the aim is that small variations in input data should not lead to changes in the prediction (Szegedy et al., 2013). However, it remains to be seen in what sense ML models can really be considered as experimental procedures, rather than solely being methods for data analysis. Moreover, adversarial robustness in ML raises some conceptual issues that are not captured by Karaca's account.

The currently most pronounced account of RA in CS, with a view on high energy physics, was developed by Boge (2021). Of particular interest in the context of ML may be his notion of 'inverse parametric robustness', referring to the insensitivity of parameter values to varying deployment conditions. While robustness of parameter values to changing data is often attained in ideal settings, e.g., via maximum likelihood methods or Bayesian updating (Lavin et al., 2021), this is often not the case in practice for more complex simulation models, e.g. in high energy physics (Boge, 2021). There, parameter values must be tailored to specific contexts, making it difficult to assign meaning to them. In this respect, complex simulation models are similar to ML—in ML, too, parameter values are highly sensitive to the deployment context and the learning setup (Quinonero-Candela et al., 2008).[6] Our framework may provide a bridge to concisely discuss such connections between ML models and simulation models.

Nevertheless, while there is much to gain from engaging with the RA literature on CS, there are some structural barriers regarding the extent that the relevant accounts can serve as a blueprint for a theory of robustness in ML. On the one hand, many accounts of RA in CS are tailored to the demands of very specific scientific domains. With regard to robustness in ML, on the other hand, we are dealing with various target systems and application scenarios. Moreover, some robustness sub-types are specific to ML—most pertinently 'adversarial robustness' or 'robustness of ML model explanations'. As a

[6] In ML, even within a context, there are a variety of parameter values that induce models with similar performance—the so-called 'Rashomon effect' (Fisher et al., 2019; Müller et al., 2023).

result, in the next section, we first take a step back and develop a framework that allows us to conceptualize robustness issues in ML in a systematic way.

3 A unified account of robustness in machine learning

3.1 Towards a definition of robustness in ML

This section provides a definition of robustness in ML—with the aim to discuss robustness research in ML more concisely.[7]

3.1.1 Robustness is a multi-place concept

In common parlance, robustness is a predicate that applies to a single entity. For instance, we might consider a person to be robust if her immune system is resilient. Similarly, a coffee machine gets called robust, if it is generally reliable. Call the entity that is robust the *robustness target*. The intuition underlying single-place notions of robustness is that that the robustness target works stable across a wide range of possible situations. However, leaving the range of situations unspecified makes it hard to operationalize the concept. In order to get a better handle on what it means for a target to be robust, it needs to be specified in relation to which other entity the robustness target remains stable—hereinafter referred to as the *robustness modifier*.[8] Robustness is therefore a multi-place concept.

3.1.2 Robustness is a causal concept

The key idea behind robustness is that the robustness target remains stable *under changes* in the robustness modifier. *Changes* is causal terminology. That is, we are concerned with two causally related entities, the target and the modifier. Presupposing an interventionist framework of causality (Woodward, 2005), a target is robust if it remains stable in light of potential manipulations to the modifier. Note that the manipulations do not have to be executed in reality: for the target to be robust, it is sufficient that the target would remain stable if they were executed.

It might be argued against this causal view of robustness in ML that the modifier and the target can be in a purely functional relationship. However, two things should be noted here: First, models, also the ones describing functional relationships, are generally implemented in physical systems (Boge, 2019). Our causal view highlights that, in ML, we are experimenting on physical implementations, which is why we do not conduct a purely formal analysis of functions (Simon, 1995). Second, even purely functional relationships can be investigated causally. If there are no causal

[7] Our framework is designed with ML robustness in mind, however, we think it could also partially be useful to analyze robustness in other (scientific) domains.

[8] While the notion of a robustness modifier is implicitly present in many papers in the philosophical robustness debate, such as Boge (2021), Schupbach (2018) and Weisberg (2006), we think it is important to make it explicit.

dependencies beyond the functional dependency between modifier and target, a causal description is still possible, even though it means taking a sledgehammer to crack a nut. However, if there are causal dependencies, a causal analysis of robustness is the only one that makes sense.

3.1.3 Robustness is limited in domain

A robustness target is rarely ever stable with respect to all possible interventions on the modifier (Boge, 2021; Schupbach, 2018; Wimsatt, 1981). If that is the case, then the modifier is not a cause of the target and robustness is trivially given. Robustness becomes a concern when the modifier causally impacts the target, but the target remains stable with respect to certain interventions on the modifier. The changes to a modifier will be referred to as its *domain*. The more significant the changes in the domain, the more difficult satisfying robustness becomes.

3.1.4 Robustness has a tolerance level

There are few, if any, applications in which a robustness target must be perfectly stable (Boge, 2021; Karaca, 2022). Rather, it is often sufficient for the robustness target to remain stable to a certain degree against changes within the modifier domain. We call the degree to which we demand target robustness the *target tolerance*. If the target tolerance is high, then robustness is less demanding and *vice versa*. What constitutes the right target tolerance depends on the application context and the assurances needed. For example, the target tolerance for an ML model designed to support clinical decision-making is considerably lower than for an ML model that has been designed to detect emails from fraudulent academic publishers.

3.1.5 A definition of robustness

Based on these considerations, we can define robustness as a causal, multi-place concept with limited domain and a specified tolerance level.[9]

Definition 1 The robustness target is said to be robust to the robustness modifier if relevant interventions in the modifier, as specified by the robustness domain, do not lead to greater changes in the target than specified by the target tolerance.

The idea expressed here is that the target must remain relatively stable with respect to specific changes on the modifier. Now, the task is to show how our framework is

[9] This definition can also be expressed mathematically using the do-operator from the causality literature (Pearl, 2009). This formalization makes the notion precise and allows its operationalization in different applications. Let $\boldsymbol{T}$ and $\boldsymbol{M}$ be random variables describing the robustness target and the robustness modifier respectively, $\mathcal{D}_{\boldsymbol{M}} \subseteq \mathcal{M}$ the domain, and $\alpha_{\boldsymbol{T}} \in \mathrm{R}_{\geq 0}$ the target tolerance. We say that $\boldsymbol{T}$ is robust with respect to $\boldsymbol{M}$ in domain $\mathcal{D}_{\boldsymbol{M}}$ with target tolerance $\alpha_{\boldsymbol{T}}$ if and only if

$$\text{for all } m \in \mathcal{D}_{\boldsymbol{M}} \text{ holds } d_{\mathcal{T}}(\boldsymbol{T}, \boldsymbol{T}_{|do(\boldsymbol{M})=m}) \leq \alpha_{\boldsymbol{T}}.$$

where $d_{\mathcal{T}}$ describes a distance function on the range of target $\boldsymbol{T}$.

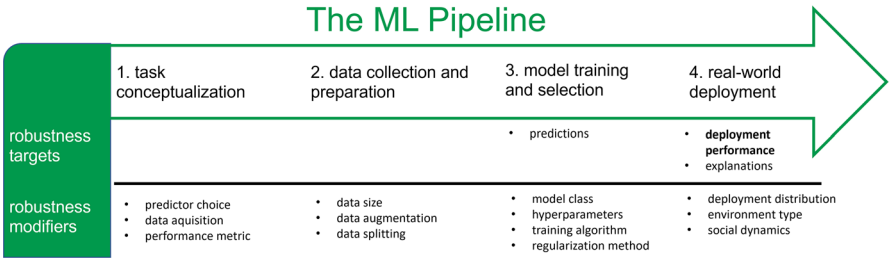

Fig. 1 The ML pipeline. It contains all relevant steps, ranging from problem conceptualization to model deployment in a real-world environment

able to capture the peculiarities in ML. What are possible robustness targets, modifiers, domains, and target tolerances? How can this framework capture current robustness discussions in ML? A precondition for answering these questions is to take a look at the ML pipeline.

3.2 The ML pipeline

The pipeline shown in Figure 1 encompasses the different steps in building and deploying supervised ML models.

3.2.1 Task conceptualization

Every supervised ML problem starts with task conceptualization. Here, developers must address fundamental questions such as: What is the *prediction target* $\boldsymbol{Y}$? What are the *predictors* $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$? From what *measurement procedure* is the (labeled) data obtained? With what *loss function* L should predictive success be measured?

3.2.2 Data collection and preparation

Once the fundamental modeling questions have been settled, ML practitioners move on to the data. Using the selected measurement procedure, they are able to draw [10] a number of data-points $(x^{(i)}, y^{(i)})_{i=1,\ldots,n}$ from $\boldsymbol{X}, \boldsymbol{Y}$, where we call n the *size* of the data. Provided that ML practitioners have additional information about the data domain, e.g., whether a certain type of noise on the data or simple transformations like mirroring or rotating images should (not) impact the prediction, they can additionally augment the data; we call the original data set together with the augmented data the *augmented data-set*. The next important step is to split this augmented data-set into three junks, the *training data*, the *validation data* and *the test data*: the training data is for training individual models; the validation data is for fitting hyper-parameters (see below for an explanation); finally, the test data is held out of the training process to make a precise estimate of the model's expected prediction error. These three data sets

[10] In supervised learning, we often assume that this data is drawn independent and identically distributed (i.i.d.).

must be randomly sampled from the augmented data-set[11], be disjoint and together form the entire data-set.

3.2.3 Model training and selection

As a next step, the modeler is searching for a good prediction model. Here, she must select an appropriate *class of models* (say tree ensembles, or a certain type of neural network), a *training algorithm* (e.g., random forest algorithm for tree ensembles, and stochastic gradient descent for neural nets), and then decide on a range of *hyper-parameters*. Hyper-parameters are important to set up the optimization problem, as they may restrict the model class or guide the model search. Examples of hyper-parameters in ML are the splitting criterion in random forests or the learning rate in stochastic gradient descent. By applying the training algorithm with specific hyper-parameters to the training data, the modeler obtains a model from the model class with low *training error*. The training error here describes the average prediction error on the training data relative to a certain notion of loss/error.

To avoid over-fitting of the model to the training data, the modeler selects the hyper-parameters based on the so-called *validation error*. Again, the validation error is estimated by the average loss, but on the validation data set rather than the training data. Importantly, the validation data set is generally not directly part of what the training algorithm *sees*. Typically, this search for *good models* is repeated for other model classes and hyper-parameter settings, leading to a set of different models with similar low validation error. The model that is finally selected for deployment is usually the one with the lowest average *test error* (also called the empirical risk).[12] The test error gives a good estimation of the *i.i.d. generalization error*, describing the true expected prediction error of the model under the assumption that the data we consider is distributed similarly to the data we observed so far.

3.2.4 Real-world deployment

Eventually, the ML developer can deploy the ML model in the real-world on the task she designed it to perform. Similar to the model training and selection process, the model is again shown data. However, unlike the previous phase, the data is processed live, often without human supervision or pre-processing. The context under which the data is taken may change and the model's predictions have practical consequences. We say, the data is sampled from the *deployment distribution*. Crucially, the deployment distribution can change over time and can be affected by predictions of the deployed ML model itself.

For economic, moral, or legal reasons, it may be necessary to explain algorithmic decisions to end-users (Vredenburgh, 2022). Consequently, these explanations may be deemed an essential part of the ML deployment and may even affect the data which our model is exposed to (König et al., 2022).

[11] Where the proportion/size of the three data-sets is again a choice the modeler has to make.

[12] The test error is defined just like the training-error, but is evaluated by using the test data rather than the training data.

In the following section, we discuss relevant robustness targets and modifiers in current robustness research. The section is not intended to provide a comprehensive picture of the research landscape on robustness in ML, but to systematize different robustness sub-types through our conceptual framework.

3.3 Robustness targets and modifiers in ML

In principle, anything in the pipeline can serve as a robustness target or a modifier. However, the modifiers must precede the targets in the pipeline, otherwise intervening on the modifiers would never affect the target. Task conceptualization precedes all conceivable targets and therefore contains universally relevant modifiers. However, it is exceedingly difficult to evaluate the influence of these modifiers. Since there is little to no research on modifiers related to task conceptualization, they are not discussed here.

In this section, our emphasis lies on the most important robustness target in current ML research—the *deployment performance*. At the same time, robustness research in ML is not exhausted by a mere focus on deployment robustness. Therefore, we also briefly discuss 'ML model predictions', and 'ML model explanations' as additional robustness targets.

To illustrate the distinct challenges for each of these robustness targets, we use the case-study (with some variations) of an ML model that predicts positive cases of COVID-19 from radio-graphic images (DeGrave et al., 2021). The example is suitable for our purposes because it is paradigmatic of an application that is subject to various robustness concerns and where a failure of robustness can have serious consequences for both, model authorities and data subjects.

3.3.1 When the target is the deployment performance

Since prediction is the main purpose of supervised learning, the *deployment performance* is the most crucial robustness target. This is closely tied to the fact that performance guarantees of the model during training are not transferable to novel settings (Koh et al., 2021). The most established modifier here is the *deployment distribution*; if the model's performance is highly sensitive to expected changes in the deployment distribution, this should be deemed as a red-flag for utilizing the model—especially if the resulting epistemic risks (Douglas, 2009) culminate in financial, legal, or health concerns. Performance robustness checks to relevant modifications in the deployment distribution are therefore integral for every model audit.

The deployment distribution itself is not the only relevant modifier in this context. Other modifiers, which can be found in Figure 1, are the task conceptualization (e.g., the choice of predictors or the loss function), data collection and preparation (e.g., data augmentation or data size), or the model training and evaluation step (e.g., the training algorithm or hyper-parameter choices). These robustness modifiers are usually studied with respect to a fixed deployment distribution scenario. For instance, we might want to examine the impact of data augmentation on the deployment performance, under the assumption that the model is used in an adversarial environment.

It is important to distinguish between *testing robustness* and *improving robustness*. When testing robustness, the task is to investigate whether an (fixed) ML model meets robustness standards within an *auditing process*. By contrast, when aiming to improve robustness, the focus is on *amelioration strategies*—which are methods that (help to) generate models, performing strongly under deployment conditions. In our framework, amelioration strategies can be described as modeling choices that improve the performance compared to default modeling choices, conditional to a given deployment distribution.

To highlight the difference between testing robustness in an audit and improving robustness with amelioration strategies, consider our COVID-19 example. When we evaluate our ML model on data from different deployment distributions (e.g., for data stemming from other X-ray devices) we are testing robustness. A positive audit (say, the deployment performance is similarly high across data from different X-ray devices) increases our confidence in the model's clinical applicability. However, provided that the deployment performance is not robust to the relevant distribution shifts, we must train a new ML model by leveraging amelioration strategies, designed to improve model performance for these scenarios. In the COVID case, we may know from experimental ML research that the predictive performance on a range of X-ray devices increases for data augmentation techniques that change the background in X-ray images. Thus, it is advisable to apply these techniques if we assume that the model will be shown images from such X-ray scanners.

As a means for identifying which robustness phenomena are at stake for a given ML application, the following questions can provide guidance: Is the deployment distribution similar to the training distribution? Does the deployment distribution naturally change over time? Does the environment adapt strategically to the model, to the detriment of the predictive performance (we call such environments *adversarial*)? Does the involvement of the predictive model itself affect the environment?

3.3.2 Robustness in the classical i.i.d. setting

While most robustness issues arise from a mismatch between the training- and the deployment distribution, robustness can also be a concern when training and deployment conditions are identical. In such cases, robustness researchers are mainly concerned with amelioration strategies that ensure strong performing models. Showing that the (high-dimensional) training and the deployment distribution satisfy the i.i.d. assumption is possible but requires vast data sets and computational resources to test with classical statistical testing methods (Lehmann et al., 2005). Thus, the i.i.d. assumption in ML is mainly justified with domain knowledge.[13] Ensuring robust performance in an i.i.d. setting is what classical ML research and statistical learning theory are all about (Bishop & Nasrabadi, 2006).

ML models can fail to i.i.d. generalize for various reasons. For instance, a model may have both, low training and low test error, but once it is deployed in a similar setting, its predictive performance drops—that means, the performance is not robust

[13] As an aside, scientific disciplines like structural biology may use large benchmark data-sets, where the data are highly standardized, satisfying the i.i.d. assumption in turn by design.

with respect to modifications in the test data sampling. This problem can occur because the data-set used for training and testing is not a statistically representative sample of the underlying data distribution. In consequence, the model learned to exploit features that are only predictive within the sample but not within the whole population (Althnian et al., 2021).

This becomes evident in the case of *shortcut learning*, where the model relies on spurious correlations in its predictions, rather than the intended features (Geirhos et al., 2020): in the example of COVID-19 prediction, imagine a case where most positive COVID cases in the training data were females, which is why the model takes the form of the pelvic to be predictive of COVID. Or suppose that a model relies on hand-written annotations on X-ray images by clinicians, which are not available under deployment conditions. Shortcut learning resembles surface learning of students in classrooms, relying on simple decision-rules in order to pass an exam (Geirhos et al., 2020). The problem of shortcuts is that they might go unnoticed in training and only occur in deployment scenarios because external validation data may not be available. This is a particularly pressing issue for rare/novel diseases. In order to prevent shortcut learning, an amelioration strategy is to remove confounding information from the data (DeGrave et al., 2021) and ensure that the data is representative of the underlying distribution.

There are several amelioration strategies for the goal of achieving i.i.d. generalization—acting on various modifiers within the ML pipeline. They all can be subsumed under the term *regularization* (Goodfellow et al., 2016). The simplest regularization technique is *data augmentation*, in which existing data instances are transformed (e.g., by flipping directions or changing the contrast in X-rays) and added to the training data (Rebuffi et al., 2021). Data augmentation increases the data size, which is a key factor for successful i.i.d. generalization (Geirhos et al., 2021). Other regularization techniques focus on the learning process rather than the data. (Li et al., 2020; Srivastava et al., 2014; Tibshirani, 1996).

3.3.3 Robustness to natural distribution shifts

The deployment distribution rarely ever remains the same as in training conditions. *Natural distribution shifts* can result from changing environmental conditions that lead to a mismatch between the deployment and the training distribution (Xie et al., 2019). Applied to the ML model used for detecting COVID-19, different lighting in the room, new cameras (with a different resolution), or changes in the patient demographics can all induce distribution shifts (Finlayson et al., 2021; Hendrycks & Dietterich, 2019; Liu et al., 2020). Time itself may even induce natural distribution shifts as it impacts societal norms, data-subjects' behavior, and large-scale environmental conditions. Ultimately, we must be aware that we are predicting future events based on data from the past, which can raise both epistemic and ethical concerns (Hardt & Kim, 2022).

One example of a natural distribution shift is a so-called *label-shift* (Lipton et al., 2018). It describes a case where the deployment distribution of the prediction target is different to its training distribution. In the COVID case, suppose that we try to predict whether a person has COVID-19, based on a set of symptoms, such as dry cough, fever, or sense of smell. For this purpose, we rely on data available from the

year 2020. However, by the end of 2022, the virus has mutated several times: features that have been clear signs of COVID in 2020, such as dry cough and a loss of smell are not as prevalent for COVID variants in 2022 (Whitaker et al., 2022). Different approaches allow to ameliorate models against label-shifts, depending on whether only the distribution of labels or also of the covariates changes (Garg et al., 2020).

Amelioration strategies to prepare for natural distribution shifts act again on various steps in the ML pipeline and include adjusting the model architecture (Li et al., 2020), using tailored data augmentation (Huang et al., 2018), modifying the optimization technique (Madry et al., 2017; Yang et al., 2020), or retraining the model (Finlayson et al., 2021).

3.3.4 Adversarial robustness

Adversarial robustness subsumes cases where the deployment environment is an adversarial actor (Huang et al., 2011). This means that the deployment distribution changes such that the ML model produces erroneous predictions. In the current literature, adversarial distribution shifts can be described as modifications to the original data distribution by adding imperceptible but intelligently designed changes/noise (Serban et al., 2020; Yuan et al., 2019).[14]

Several papers have shown that standardly trained ML models are not robust to such modifications in the deployment distribution, called *adversarial attacks* (Goodfellow et al., 2014; Szegedy et al., 2013); these attacks can be efficiently executed in real-world scenarios (Athalye et al., 2018; Song et al., 2018) and performed even if the attacker has only application programming interface (API) access to the model—as these attacks transfer between different models (Papernot et al., 2016). As an example of an adversarial attack in the medical domain, consider a case where humanly imperceptible noise is added to medical images to fool the ML model into classifying moles as malignant skin tumors, while reporting high confidence (Finlayson et al., 2019).

Fundamental research on the causes of ML models' vulnerability to adversarial attacks led to various amelioration strategies: in the context of adversarial attacks, amelioration strategies are usually called *defenses* (Goodfellow et al., 2014; Tanay & Griffin, 2016).[15] However, common defenses based on data augmentation (adversarial training), architectural defenses, or anomaly detection are only partially successful: they just tend to work for certain adversarial attacks (Serban et al., 2020; Shafahi et al., 2019; Zantedeschi et al., 2017). The effectiveness of such defense strategies is therefore subject to controversies, particularly since they might (negatively) impact the i.i.d. performance by constraining the amount of available predictive features (Tsipras et al., 2018; Yang et al., 2020).

One interpretation about the *nature* of adversarial attacks dominates in recent research. According to this view, the vulnerability of ML models to adversarial attacks

[14] Adversarial robustness also includes *data-poisoning*—e.g., the modification of the training data prior to the training process with the goal of creating weak-spots in the model (Biggio et al., 2012). Since we are concerned with modifications in the deployment and not in the training distribution, we focus on adversarial robustness in the above sense.

[15] Similar to the defenses against adversarial attacks, there are also defenses against data poisoning (Steinhardt et al., 2017).

results from the very nature of association-based supervised ML (Buckner, 2020; Ilyas et al., 2019); the models learn to exploit all patterns inherent in the data that contain predictive information, including such patterns that are inscrutable to human cognition or only associative but not causal for the prediction target (Freiesleben, 2022). The reliance on these non-causal, humanly inscrutable patterns is exploited in adversarial attacks, leading to the vulnerability of ML models to such changes. This view contributes to a revised understanding of what is required to cope with adversarial attacks: The problem becomes to separate legitimate predictive features from non-legitimate humanly inscrutable predictive features (Buckner, 2020), and causes from mere associations (Schölkopf, 2022); both tasks are tremendously hard and cannot be readily used to design novel defenses.

3.3.5 Robustness to model-induced distribution shifts

The model's predictions themselves can lead to distribution shifts by causing changes in people's behavior. This phenomenon is called *performative prediction* (Perdomo et al., 2020). The underlying problem is common in social science experiments: in virtue of their being studied upon, research subjects may (unconsciously or deliberately) adjust their behavior, which then impacts the validity of the generated data (Jiménez-Buedo, 2021).[16] To illustrate this with a variation of our COVID prediction model: assuming that the model is used to inform triage decisions (e.g., by estimating the risk of severe disease progression), certain patient groups (high-risk patients or people distrusting algorithms) may refuse to visit a given hospital. In turn, the model might be used for demographics other than those for which it was trained.

To preserve high model-performance, the most common amelioration strategy is to retrain the model on the new data; but again, the retrained model enters the performativity-cycle. Interestingly, there exist equilibrium states where the performance is high and predictions do not result in further distribution shifts (Brown et al., 2022; Perdomo et al., 2020). That being said, these equilibrium states are bound to very demanding conditions, which is why the results also permit a negative interpretation.

One special kind of performative prediction that is well-studied in ML is *strategic classification*. It describes distribution shifts that occur because individuals strategically optimize their behavior to reach a desired classification outcome—e.g., by making small financial transactions to boost one's credit score (Hardt et al., 2016). Preventing strategic behavior that leads the model to mis-classifications (also called gaming), while incentivizing agents to achieve their desired outcome through adequate means, proves difficult for developers and policymakers, not the least because it involves a challenging causal inference problem (Miller et al., 2020).

3.3.6 Robustness targets beyond performance

Having discussed robustness phenomena related to predictive performance, we now turn to other robustness targets:

[16] In the philosophy of social science, reactivity is the more common term than performativity.

3.3.7 When the target is ML model predictions

Predictions of an ML model f are another robustness target, with the feature values of the data being the most central modifier. Here, the goal is to determine whether the ML model makes its predictions based on meaningful features. To give an example from clinical medicine, Tomašev et al. (2019) developed an ML model that predicts the risk of acute kidney failure for patients up to 48 hours in advance. In the course of the model evaluation, the investigators manipulated the feature weights of diagnostic markers to assess how it impacts the predicted output. This is to ensure that the model tracks clinically meaningful information, increasing confidence in its clinical applicability in turn. Model-agnostic IML[17] is another strategy to detect biases in ML models. For instance, individual conditional expectation (ICE) curves describe the sensitivity of individual predictions to changes in a single feature (Goldstein et al., 2015).

Aside from providing sanity checks, *procedural algorithmic fairness* can be another objective regarding the robustness of model predictions. The underlying idea here is that the model makes accurate predictions across different demographic groups for the *right* reasons (Barocas et al., 2017; Schwöbel & Remmers, 2022). Pragmatically, for many instances of consequential decision-making, this means that the predictions should not be affected by sensitive attributes, such as gender or ethnicity (Grgić-Hlača et al., 2018; Kusner et al., 2017; Morseet al., 2021). Several amelioration strategies have been proposed to enable procedurally fair ML models, relying on data augmentation (Sharma et al., 2020), architectural adjustments (Du et al., 2020), and feature selection (Dwork et al., 2012; Gajane & Pechenizkiy, 2017; Grgić-Hlača et al., 2018). Intriguingly, robustness and fairness are increasingly thought together in model evaluation (Li et al., 2021; Xu et al., 2021), which is why it has been argued that they converge (Lee et al., 2021).

3.3.8 When the target is ML explanations

Explanations are increasingly becoming an important aspect of the ML pipeline when it comes to deployment in high-stakes scenarios. After all, judges, clinicians, and policymakers need to understand and justify why consequential decisions are being made. As a result, ML explanations have established themselves as an important robustness target. The primary concern here is to which extent the explanations provided by various Interpretable Machine Learning (IML) methods faithfully represent the *inner logic* of the ML model (Rudin, 2019). Several ML research papers have demonstrated that standard IML methods provide non-robust explanations with respect to modifications in the trained ML model, hyperparameter choices of the IML technique, or changes to the input: Popular IML techniques such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), and counterfactual explanations (Wachter et al., 2017) can all be tricked to provide any desired explanation by modifying the model in areas without

[17] The name model-agnostic stems from the fact that the model is seen as an input-output mapping, ignoring the specific model structure (Scholbeck et al., 2019).

data support (Lakkaraju & Bastani, 2020; Slack et al., 2020, 2021);[18] In a similar manner, methods that aim to highlight the model's attention in a decision (called *saliency based techniques*) are highly unstable in light of constant or small noisy changes in images (Ghorbani et al., 2019; Kindermans et al., 2019); Surprisingly, saliency maps have shown to be robust to the specifics of the ML model, which gives an example of a robustness statement that must be interpreted negatively (Adebayo et al., 2018); Lastly, feature importance scores vary across equally well performing models Fisher et al. (2019).

Ensuring the robustness of explanations becomes particularly challenging for *recourse recommendations*, whose purpose is to inform end-users of potential actions they can take to achieve a desired outcome, such as getting a loan request accepted (Hancox-Li, 2020; Venkatasubramanian & Alfano, 2020). In this vein, Upadhyay et al. (2021) studies robustness failures in recourse explanations due to model shifts, whereas König et al. (2022) show that recourse explanations themselves may induce distribution shifts in the deployment distribution—in virtue of which recourse recommendations often fail to provide reliable action guidance. This is because recourse recommendations provoke end-users to change their behavior. To counteract this *performativity* of recourse recommendations, König et al. (2022) develop an amelioration strategy whose basic idea is to constrain the given explanations based on causal knowledge.

3.4 Modifier domain and target tolerances in ML

Modifier domains describe a range of situations that the modifier might be in. Depending on the modifier, this can be a set of alternative data splits, regularization strategies, or deployment distributions. The choice of a good modifier domain is up to the user and is always context dependent. What modification might occur? What is the worst-case scenario for which we need to provide guarantees? Universal modifier domains generally make little sense, since there always exist some deployment distributions where the model performance drops—as has been demonstrated by the no-free lunch theorems (Sterkenburg & Grünwald, 2021; Wolpert, 2002).

Just like the modifier domains, the target tolerance is context specific. It depends on the degree of robustness needed, the distance function, and also the scaling of features/performances. In light of this, the modifier domain, distance function, and target tolerance can be seen as the fine-tuning parameters in the study of robustness.

4 The irreducibility of robustness

In Sect. 2.1, we raised the possibility that robustness in ML may be reducible to other concepts. In this section, we now refute this idea by arguing that robustness is neither reducible to, nor identical with, other central concepts in ML, such as extrapolation or generalization. Instead, our claim is that robustness is a complementary epistemic

[18] Adapting the selection of data points to generate the explanation may defend against some of these attacks but not all of them (Vreš & Šikonja, 2021).

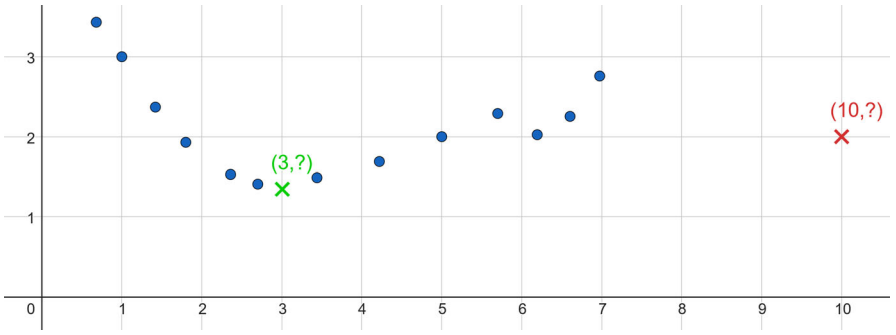

Fig. 2 Interpolation vs extrapolation. The blue dots describe the data along predictor x and target y. For value $x = 3$ (green), the correct target value y is unknown, but x lies close to values for which we know y—so we must interpolate to predict y for $x = 3$. For value $x = 10$ (red), the correct target value is also unknown, but it lies outside of the range of the data we have—thus we must extrapolate to predict y for $x = 10$

concept to uncertainty quantification regarding the deployment of ML models under real-world conditions.

4.1 Robustness is not reducible to extrapolation

Extrapolation describes the prediction of data that lies geometrically outside of the manifold, spanned by the training data (Brezinski & Zaglia, 2013). Usually, extrapolation is contrasted with *interpolation*, describing the prediction of data that lies on this manifold (Bishop & Nasrabadi, 2006; Goodfellow et al., 2016; Hasson et al., 2020). However, it is difficult to draw a clear line between interpolation and extrapolation in practice, because how exactly the manifold is spanned by the training data is controversial: is it only the data-set itself, its linear hull, its convex span, or some completely different geometric closure of a set (see Figure 2 for an intuition)?

What is less controversial,[19] is that current ML models are designed for interpolation, rather than extrapolation (Barnard & Wessels, 1992; Haley & Soloway, 1992). Extrapolation only works if we incorporate a strong inductive bias in learning: that is, if we encode knowledge about the data generating process into the learning procedure (Gordon & Desjardins, 1995; Mitchell, 1980). It is debatable whether deep neural nets can exhibit such a strong inductive bias, as it has been shown that they can even fit random noise perfectly (Zhang et al., 2021).

It seems as if extrapolation and robustness handle similar problems, provided that both are concerned with reliable predictions on data that might lie outside of the training data. However, there is an important difference: in the case of extrapolation, we assume that the prediction task has not really changed. What differs, instead, is the data; the data either stems from the same distribution as the training data but is not represented in it. Alternatively, it results from a change in the distribution of predictors $\mathbb{P}(\boldsymbol{X})$ (this is often referred to as a covariate shift). What must remain constant in extrapolation

[19] (Balestriero et al., 2021) see the strengths of ML in extrapolation. However, they use the convex hull for their notion of manifold, which can be argued is too narrow.

is the relationship between the predictors and the target i.e. $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$. For robustness, this is not necessary (think of label-shifts). Thus, at most, extrapolation is a sub-type of robustness but not vice versa.

Extrapolation is often studied in the context of causal inference, namely the extrapolation of causal effects to a target system under real-world interventions (Khosrowi, 2022; Reiss, 2019). This causal notion of extrapolation comes closer to our notion of robustness: real-world interventions lead to distribution shifts, and conversely, distribution shifts can always be described as interventions in the data-generating causal mechanism (Pearl, 2009). However, this causal strand of research presumes a very different conceptual framework, focusing on causal assumptions about the relation between the target system and the original model. Our account, by contrast, only makes assumptions about the deployment distribution.

4.2 Robustness goes beyond (i.i.d) generalization

A model *i.i.d. generalizes* if it has low error on unseen test data from the same distribution as the training data. One reason why ML models might fail to i.i.d. generalize is that they fail to interpolate the data due to over- or under-fitting (Bishop & Nasrabadi, 2006; Hastie et al., 2009). Alternatively, the model may fail to extrapolate due to an unsuitable inductive bias, or because it has been trained with too little data.

I.i.d. generalization does not capture robustness because it does not deal with predictive performance in the presence of distribution shifts. Instead, i.i.d. generalization provides performance assurances in a static environment. For performance robustness instead, everything revolves around changes in the environment and their impact on model performance.

Nevertheless, i.i.d. generalization is a necessary, though not sufficient condition for robustness. If an ML model does not generalize well under i.i.d. conditions, robustness is the least of its problems. Robustness, by our definition, is a relative concept that refers to the initial state of the model under training conditions. For example, if the initial model fits the training data poorly, it will most likely also perform poorly under distribution shifts. Consequently, even though robustness as a concept is generally applicable, an analysis of robustness in ML models is only meaningful if a sufficient level of i.i.d. generalization has been achieved.

One might argue that generalization is really not about the model's performance in an i.i.d. setting, but rather in an out of distribution (o.o.d.) setting (Geirhos et al., 2020). O.o.d. is usually contrasted with i.i.d., highlighting that the samples may be drawn from a different distribution (e.g., due to distribution shift) or they are not sampled independently. Even though intuitively, o.o.d. generalization is similar to our notion of robustness, there are good reasons to prefer robustness over o.o.d. generalization as a concept: Firstly, o.o.d. generalization as a formal predicate lacks a proper operationalization of its success conditions. It remains unspecified to which distribution exactly the model should generalize, how strict the notion for marginal performance drops is, or if o.o.d. generalization requires i.i.d. generalization. Precisely these success conditions have been provided for robustness in this paper. Secondly, o.o.d. generalization puts the focus exclusively on the deployment performance, whereas robustness is the

more powerful meta-concept that can be applied to any target, including predictions or explanations. Finally, unlike o.o.d. generalization, robustness in ML is an inherently causal concept. It therefore goes beyond model audits, by enabling control and informing amelioration strategies.

4.3 Robustness complements uncertainty quantification

Building on our conceptual framework for robustness, it can be argued that the objective of *uncertainty quantification* in ML is to estimate the sensitivity of a target state with respect to an indeterminate modifier scenario. In plain words, how do uncertainties in the modifiers within the ML pipeline propagate forward to the relevant target?

There can be different kinds of uncertainties involved in ML, dependent on the modifier: (i) uncertainties can arise in our data, e.g., due to flawed instruments (Abdar et al., 2021; Hüllermeier & Waegeman, 2021); (ii) there can be uncertainties in the learning process, since many optimizers and models contain stochasticity (Hennig et al., 2015; Molnar et al., 2021; Nadeau & Bengio, 1999); (iii) or, there can be uncertainties about the deployment environment faced (Abdar et al., 2021; Koh et al., 2021).

Interestingly, the first two of these types of uncertainties have counterparts in the philosophical debate on CS, by way of uncertainties related to measurement and uncertainties due to computation (Boge, 2021; Parker, 2017; Tal, 2012). And the parallels between uncertainty in ML and CS go even further. For instance, the concepts of aleatoric (non-reducible) and epistemic (reducible) uncertainty in ML (Hüllermeier & Waegeman, 2021) closely match the notions of statistical and systematic uncertainty by Staley (2020). Similarly, different sources of epistemic uncertainty, such as limited computational resources or imperfect modeling, closely matches what Parker (2017) calls dynamical model error and numerical error.

Relevant uncertainty targets in ML are the same as for robustness targets: predictions, performance, and explanations—and many papers in the field evaluate uncertainty for these targets (Hüllermeier & Waegeman, 2021; Molnar et al., 2021; Nadeau & Bengio, 1999). We see that robustness and uncertainty quantification are both powerful meta-concepts in ML for assessing the reliability of model properties.[20] Given the strong similarity between uncertainty quantification and robustness in ML, is there a way to delineate both notions?

There are indeed clear differences between the two concepts in ML. First, in uncertainty quantification, the modifier states are weighted according to their likeliness. By contrast, in robustness, all states that are within the modifier domain are on equal footing. While a single outlier scenario within the modifier domain may break robustness, the uncertainty of the target state may remain low because that outlier scenario is considered unlikely. Second, robustness only considers scenarios within the modifier domain, whereas uncertainty quantification considers all possible scenarios—although

[20] Indeed, they can be even applied to each other; one can study the robustness of uncertainty quantification (Hein et al., 2019; Kristiadi et al., 2020), or quantify the uncertainty of performance robustness (Nadeau & Bengio, 1999).

it may not assign positive weight to all of them. Third, uncertainty quantification estimates the uncertainty in the target state, however large it may be. Robustness, on the other hand, is a binary predicate that describes whether the target state does not change more than specified by the target tolerance.

More fundamentally, robustness as a property encodes causal knowledge about the object of interest. That is, it tells model-developers that the robustness target remains stable (given a certain tolerance level), compared to an initial state under certain interventions. Even though uncertainty encodes similar knowledge, the two can be seen as complementary notions when it comes to guiding actions. Robustness assures that even in a worst-case scenario, the robustness target does not change more than specified by the tolerance level. Contrastingly, even if it would have devastating effects, uncertainty quantification may give us optimistic estimates for the expected target state, provided that the worst-case scenario is unlikely to occur.

Although both robustness and uncertainty quantification are meta-concepts that can help to analyze the sensitivity of any target, we find that robustness is most beneficial when it comes to the overall model evaluation, whereas uncertainty quantification may be more crucial for individual predictions. Uncertainty quantification is particularly important when the model is utilized under conditions where assurances for the overall performance are insufficient; think of cases where an ML model is used as a support tool to guide consequential decision-making. Here, communicating the model's uncertainty enables policymakers to assess the model's confidence for single instances. If the reported uncertainty is low, the human decision-maker may defer to the model, whereas if the uncertainty is high (by whatever predefined standard), decision-makers should abstain from giving much weight to the algorithmic predicted output (Kompa et al., 2021).

Unlike for individual predictions, uncertainty quantification is often less feasible for overall model evaluation: Assigning probabilities to the target state necessitates assigning probabilities to scenarios in the modifier domain. However, especially for the modifiers that are most relevant for model performance, such as learning algorithms, data splitting, or deployment distributions, such an assignment either does not make any sense or the necessary information is not available.

5 Conclusion

Whenever ML models are deployed in environments where the stakes are high, it is important to implement appropriate guardrails. However, i.i.d generalization by itself is insufficient to provide the necessary assurances, since even high performing ML models are vulnerable to changes in the deployment distribution. We believe that the concept of robustness—having distinct epistemic functions from other concepts in ML and statistics—can help to provide much needed guardrails.

The starting point of this paper was the observation that, so far, the usage of robustness in ML has either been entirely context dependent or simply left vague. The key contribution of our framework is to provide a common language for robustness research. What is meant by robustness is fully specified by a given robustness target,

modifier(-domain), and target tolerance. Through this lens, we have analyzed a variety of robustness sub-types, while also discussing possible strategies for improving robustness in ML. Lastly, we hope our work emphasizes the value of (philosophical) conceptual analysis in the fast-moving landscape of ML research.

Looking ahead, one task of future work could be to investigate the role of robustness within the wider nexus of trustworthy and reliable ML (Duede, 2022; Durán & Formanek, 2018). In scientific contexts, for example, it would be important to explore and understand the relationship between ML robustness and the reproducibility of ML models. In domains such as clinical medicine or criminal justice, again, it may not suffice to confine the analysis on model performance. Rather, the ML model must be understood as part of a socio-technical system in which the successful interplay between the ML model and human decision-makers is crucial. This interplay between model and human, in turn, may require additional guarantees, such as prospective studies on clinician-ML interaction (Genin & Grote, 2021). Overall, we hope that this paper will prove to be a useful point of reference for future research on trustworthy ML.

Acknowledgements Both authors contributed equally to the paper. We thank Tom Sterkenburg, Gunnar König, Emily Sullivan, and two anonymous reviewers for their helpful feedback on the manuscript. We acknowledge support by the Carl Zeiss Foundation (Project: Certification and Foundations of Safe Machine Learning Systems in Healthcare) and the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence "Machine Learning-New Perspectives for Science", EXC 2064, Project Number 390727645).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243–297.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems, 31*, 9505–9515.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A., & Kurdi, H. (2021). Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences, 11*(2), 796.

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G. K., Rogelj, J., Rojas, M., Sillma, J., Storelvmo, T., Thorne, P. W., Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R. P., Armour, K., & Zickfeld, K. (2021). Climate change 2021: The physical science basis. Contribution of working group 14 I to the sixth assessment report of the Intergovernmental Panel on Climate Change. Technical Summary.

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *International conference on machine learning* (pp. 284–293).

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. arXiv Preprint. arXiv:2110.09485

Barnard, E., & Wessels, L. (1992). Extrapolation and interpolation in neural network classifiers. *IEEE Control Systems Magazine, 12*(5), 50–53.

Barocas, S., Hardt, M., & Narayanan, A. (2017). *Fairness in machine learning. NIPS Tutorial, 1*, 2.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–12).

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. arXiv Preprint. arXiv:1206.6389

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Boge, F. J. (2019). Why computer simulations are not inferences, and in what sense they are experiments. *European Journal for Philosophy of Science, 9*, 1–30.

Boge, F. J. (2021). Why trust a simulation? Models, parameters, and robustness in simulation-infected experiments. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/716542

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

Brezinski, C., & Zaglia, M. R. (2013). *Extrapolation methods: theory and practice*. Elsevier.

Brown, G., Hod, S., & Kalemaj, I. (2022). Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics* (pp. 6045–6061). PMLR.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence, 2*(12), 731–736.

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

DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence, 3*(7), 610–619.

Dong, Y., Fu, Q.A., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2020). Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 321–331).

Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Drenkow, N., Sani, N., Shpitser, I., & Unberath, M. (2021). Robustness in deep learning for computer vision: Mind the gap? arXiv Preprint. arXiv:2112.00639

Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A., & Seshia, S.A. (2019). A formalization of robustness for deep neural networks. arXiv Preprint. arXiv:1903.10033

Du, M., Yang, F., Zou, N., & Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems, 36*(4), 25–34.

Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese, 200*(6), 1–20.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645–666.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science, 363*(6433), 1287–1289.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine, 385*(3), 283.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

Freiesleben, T. (2022). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines, 32*(1), 77–109.

Freiesleben, T., König, G., Molnar, C., & Tejero-Cantero, A. (2022). Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv Preprint. arXiv:2206.05487

Gajane, P. & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. arXiv Preprint. arXiv:1710.03184

Garg, S., Wu, Y., Balakrishnan, S., & Lipton, Z. (2020). A unified view of label shift estimation. *Advances in Neural Information Processing Systems, 33*, 3290–3300.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems, 34*, 23885–23899.

Genin, K., & Grote, T. (2021). Randomized controlled trials in medical AI: A methodological critique. *Philosophy of Medicine, 2*(1), 1–15.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 3681–3688).

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44–65.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT.

Goodfellow, I. J., Shlens, J., & sSzegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv Preprint. arXiv:1412.6572

Gordon, D. F., & Desjardins, M. (1995). Evaluation and selection of biases in machine learning. *Machine Learning, 20*(1), 5–22.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32).

Gueguen, M. (2020). On robustness in cosmological simulations. *Philosophy of Science, 87*(5), 1197–1208.

Haley, P. J., & Soloway, D. (1992). Extrapolation limitations of multilayer feedforward neural networks. In *[Proceedings 1992] IJCNN international joint conference on neural networks* (Vol. 4, pp. 25–30). IEEE.

Ham, Y. G., Kim, J. H., & Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. *Nature, 573*(7775), 568–572.

Hancox-Li, L. (2020). Robustness in machine learning explanations: Does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 640–647).

Hardt, M., Kim, M. P. (2022). Backward baselines: Is your model predicting the past? arXiv Preprint. arXiv:2206.11673

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science* (pp. 111–122).

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416–434.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction,* (Vol. 2). Springer.

Hein, M., Andriushchenko, M., & Bitterwolf, J. (2019). Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 41–50).

Heiri, O., Brooks, S. J., Renssen, H., Bedford, A., Hazekamp, M., Ilyashuk, B., Jeffers, E. S., Lang, B., Kirilova, E., Kuiper, S., Millet, L., Samartin, S., Toth, M., Verbruggen, F., Watson, J. E., van Asch, N., Lammertsma, E., Amon, L., Birks, H. H., & Lotter, A. F. (2014). Validation of climate model-inferred regional temperature change for late-glacial Europe. *Nature Communications, 5*(1), 1–7.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, M., Steinhardt, J., & Gilmer, J. (2021). The many faces of robustness: A critical analysis

of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8340–8349).
Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv Preprint. arXiv:1903.12261
Hennig, P., Osborne, M. A., & Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471*(2179), 20150142.
Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on security and artificial intelligence* (pp. 43–58).
Huang, Y., Würfl, T., Breininger, K., Liu, L., Lauritsch, G., & Maier, A. (2018). Some investigations on robustness of deep learning in limited angle tomography. In *International conference on medical image computing and computer-assisted intervention* (pp. 145–153). Springer.
Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*, 457–506.
Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *33rd Conference on neural information processing systems (NeurIPS 2019)* (Vol. 32).
Jiménez-Buedo, M. (2021). Reactivity in social scientific experiments: What is it and how is it different (and worse) than a placebo effect? *European Journal for Philosophy of Science, 11*(2), 1–22.
Karaca, K. (2022). Two senses of experimental robustness: Result robustness and procedure robustness. *The British Journal for the Philosophy of Science, 73*(1), 279–298.
Khosrowi, D. (2022). What's (successful) extrapolation? *Journal of Economic Methodology, 29*(2), 140–152.
Kindermans, P. J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). *The (un) reliability of saliency methods, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 267–280*. Springer.
Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., & Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning* (pp. 5637–5664). PMLR.
Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine, 4*(1), 1–6.
König, G., Freiesleben, T., & Grosse-Wentrup, M. (2022). Improvement-focused causal recourse (icr). arXiv Preprint. arXiv:2210.15709
Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience, 21*(9), 1148–1160.
Kristiadi, A., Hein, M., & Hennig, P. (2020). Being Bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning* (pp. 5436-5446). PMLR.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science, 61*(3), 541–567.
Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems* (Vol. 30).
Lakkaraju, H., & Bastani, O. (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* (pp. 79–85).
Lavin, A., Krakauer, D., Zenil, H., Gottschlich, J., Mattson, T., Brehmer, J., Anandkumar, A., Choudry, S., Rocki, K., Baydin, A.G., Prunkl, C., Paige, B., Isayev, O., Peterson, E., McMahon, P. L., Macke, J., Cranmer, K., Zhang, J., Wainwright, H., & Pfeffer, A. (2021). Simulation intelligence: Towards a new generation of scientific methods. arXiv Preprint. arXiv:2112.03235
Lee, J.G., Roh, Y., Song, H., & Whang, S. E. (2021). Machine learning robustness, fairness, and their convergence. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 4046–4047).
Lehmann, E. L., Romano, J. P., & Casella, G. (2005). *Testing statistical hypotheses* (Vol. 3). Springer.
Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–431.
Li, M., Soltanolkotabi, M., & Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics* (pp. 4313–4324). PMLR.

Li, Q., Shen, L., Guo, S., & Lai, Z. (2020). Wavelet integrated CNNs for noise-robust image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7245–7254).
Li, T., Hu, S., Beirami, A., & Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning* (pp. 6357–6368). PMLR.
Lipton, Z., Wang, Y. X., & Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning* (pp. 3122–3130). PMLR.
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.
Liu, X., Xiao, T., Si, S., Cao, Q., Kumar, S., & Hsieh, C. J. (2020). How does noise help robustness? Explanation and exploration under the neural sde framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 282–290).
Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984.
Lundberg, S. M. & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30).
Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv Preprint. arXiv:1706.06083
Miller, J., Milli, S., & Hardt, M. (2020). Strategic classification is causal modeling in disguise. In *International conference on machine learning* (pp. 6917–6926). PMLR.
Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Rutgers CS Tech Report CBM-TR-117.
Molnar, C. (2020). *Interpretable machine learning*. www.Lulu.com
Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021). Relating the partial dependence plot and permutation feature importance to the data generating process. arXiv Preprint. arXiv:2109.01433
Morse, L., Teodorescu, M. H. M., Awwad, Y., & Kane, G. C. (2021). Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics, 181*, 1083–1095.
Müller, S., Toborek, V., Beckh, K., Bauckhage, M. J. C., & Welke, P. (2023). An empirical evaluation of the Rashomon effect in explainable machine learning. arXiv Preprint. arXiv:2306.15786
Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. In *Advances in neural information processing systems* (Vol. 12).
Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins's the strategy of model building in population biology (1966). *The Quarterly Review of Biology, 68*(4), 533–546.
Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv Preprint. arXiv:1605.07277
Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600.
Parker, W. S. (2017). Computer simulation, measurement, and data assimilation. *The British Journal for the Philosophy of Science, 68*(1), 273–304.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International conference on machine learning* (pp. 7599–7609). PMLR.
Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset shift in machine learning*. MIT.
Ranjan, A., Janai, J., Geiger, A., & Black, M. J. (2019). Attacking optical flow. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2404–2413).
Rebuffi, S. A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. *Advances in Neural Information Processing Systems, 34*, 29935–29948.
Reiss, J. (2019). Against external validity. *Synthese, 196*(8), 3103–3121.
Ribeiro, M. T., Singh, S., & Guestrin, C. (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* (pp. 1135–1144).
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.
Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., & Casalicchio, G. (2019). Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 205-216). Springer.

Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl* (pp. 765–804). ACM.
Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science, 69*(1), 275–300.
Schwöbel, P., & Remmers, P. (2022). The long arc of fairness: Formalisations and ethical discourse. In *2022 ACM conference on fairness, accountability, and transparency, FAccT '22*, New York, NY, USA (pp. 2179–2188). Association for Computing Machinery.
Serban, A., Poll, E., & Visser, J. (2020). Adversarial examples on object recognition: A comprehensive survey. *ACM Computing Surveys (CSUR), 53*(3), 1–38.
Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! In *Advances in neural information processing systems* (Vol. 32).
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020). Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 358–364).
Simon, H. A. (1995). Artificial intelligence: An empirical science. *Artificial intelligence, 77*(1), 95–127.
Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems, 34*, 9391–9404.
Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 180–186).
Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., & Kohno, T. (2018). Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.
Staley, K. W. (2020). Securing the empirical value of measurement results. *The British Journal for the Philosophy of Science, 71*(1), 87–113.
Steinhardt, J., Koh, P. W. W., & Liang, P. S. (2017). Certified defenses for data poisoning attacks. In *Advances in neural information processing systems* (Vol. 30).
Sterkenburg, T. F., & Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese, 199*(3), 9979–10015.
Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science, 73*(1), 109–133.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergu, R. (2013). Intriguing properties of neural networks. arXiv Preprint. arXiv:1312.6199
Tal, E. (2012). *The epistemology of measurement: A model-based account*. University of Toronto.
Tanay, T. & Griffin, L. (2016). A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint* arXiv:1608.07690
Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems, 33*, 18583–18599.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.
Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Peterson, K., & Mohamed, S. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature, 572*(7767), 116–119.
Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. arXiv Preprint. arXiv:1805.12152
Upadhyay, S., Joshi, S., & Lakkaraju, H. (2021). Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems, 34*, 16926–16937.
Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 284–293).

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy, 30*(2), 209–229.
Vreš, D., & Šikonja, M. R. (2021). Better sampling in explanation methods can prevent dieselgate-like deception. arXiv Preprint. arXiv:2101.11702
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*, 841.
Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese, 200*(1), 1–33.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–742.
Whitaker, M., Elliott, J., Bodinier, B., Barclay, W., Ward, H., Cooke, G., Donnelly, C. A., Chadeau-Hyam, M., & Elliott, P. (2022). Variant-specific symptoms of covid-19 in a study of 1,542,510 adults in England. *Nature Communications, 13*(1), 1–10.
Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), *Scientific inquiry and the social science* (pp. 124–163). Jossey-Bass.
Wolpert, D.H. (2002). The supervised learning no-free-lunch theorems. *Soft Computing and Industry*: 25–42.
Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219–240.
Xie, R., Yu, F., Wang, J., Wang, Y., & Zhang, L. (2019). Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.
Xu, H., Liu, X., Li, Y., Jain, A., & Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning* (pp. 11492–11501). PMLR.
Yang, T., Zhu, S., & Chen, C. (2020). Gradaug: A new regularization method for deep neural networks. *Advances in Neural Information Processing Systems, 33*, 14207–14218.
Yang, Y. Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., & Chaudhuri, K. (2020). A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems, 33*, 8588–8601.
Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems, 30*(9), 2805–2824.
Zantedeschi, V., Nicolae, M. I., & Rawat, A. (2017). Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 39–49).
Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine, 15*(11), e1002683.
Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34*(2), 265–288.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115.
Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications*. Springer.