Journal of Computational Science 83 (2024) 102452

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Learnability of state spaces of physical systems is undecidable

Petr Spelda *, Vit Stritecky

Department of Security Studies, Institute of Political Studies, Faculty of Social Sciences, Charles University, U Kříže 8, 158 00, Praha 5, Czech Republic

ARTICLE INFO

Keywords:
Undecidability
Machine learning
Probably approximately correct learning
Scientific exploration
Deep neural networks

ABSTRACT

Despite an increasing role of machine learning in science, there is a lack of results on limits of empirical exploration aided by machine learning. In this paper, we construct one such limit by proving undecidability of learnability of state spaces of physical systems. We characterize state spaces as binary hypothesis classes of the computable Probably Approximately Correct learning framework. This leads to identifying the first limit for learnability of state spaces in the agnostic setting. Further, using the fact that finiteness of the combinatorial dimension of hypothesis classes is undecidable, we derive undecidability for learnability of state spaces as well. Throughout the paper, we try to connect our formal results with modern neural networks. This allows us to bring the limits close to the current practice and make a first step in connecting scientific exploration aided by machine learning with results from learning theory.

1. Introduction

Machine learning models have been integrated into methodological toolboxes of several scientific disciplines. In physical sciences, there are many applications with promising results [14,30]. Exploration in particle physics [12,13,17,32,45], materials science [18,21] or astrophysics [16,29,41] is supported by machine learning models that classify and generate observational evidence and even replace traditional simulations. These models enable access to the space of possible states, i.e., state space, of the physical system characterized by theories about the modelled phenomenon. By performing inferences about possible states of the physical system, the models provide modal information about the phenomenon. That is, information delimiting the state space of the physical system by providing answers about what is (presumably) possible with respect to the system in question.

Philosophers (of science) call such efforts modal modelling [47,51,56] and ask what makes modal inferences reliable with respect to state spaces [34,46]. Modal modelling is treated as an inferential tool used to refine or even possibly replace theories characterizing physical systems that underlie the state spaces ([47], pp. 207–208). Machine learning models used in physical sciences, serving as classifiers or generators of observational evidence, play this modal modelling role. The philosophers' worry about reliability of inferences used to access the state spaces of physical systems seems, thus, valid.

Here we show that the reliability issue cannot be ignored because it cannot be resolved. A binary classifier implementable in an artificial neural network, predicting whether observations belong to some state space, can be used to prove that learnability of state spaces is formally undecidable. This creates a significant obstacle for justifying modal inferences performed by machine learning models and, as a result, also a formal limit for uses of machine learning models in scientific exploration.

Scientific exploration can be considered a process of acquiring modal knowledge about state spaces of physical systems. If the exploration is not enhanced by machine learning models, modal knowledge is usually produced using counterfactual conditionals that describe different states of the system derived from some causal model [55]. In case artificial neural networks (ANNs) are involved, the access to modal knowledge is based also on their universal approximation property.

Universal approximation gives ANNs the ability to approximate measurable functions to arbitrary accuracies provided that the network has a sufficient representational capacity given as the number of hidden units ([27,43]; made precise by bounding ANNs' representational capacities for approximating different function classes, e.g., [4]). From the perspective of causal learning, universal approximation lacks a world model which would allow it to learn causal relations between inputs and outputs and to infer counterfactuals from the relations [40]. ANNs represent a way to learn relations between input-output pairs without causal world models. The relations are then used to perform valid inferences on new inputs.

We show that learnability of computable access to state space modalities based on universal approximation is formally undecidable (see

* Corresponding author.
E-mail address: petr.spelda@fsv.cuni.cz (P. Spelda).

https://doi.org/10.1016/j.jocs.2024.102452
Received 8 January 2024; Received in revised form 25 July 2024; Accepted 25 September 2024
Available online 26 September 2024

Theorem 1 in Section 4 for the main result). We proceed in two steps. First, we recap the fundamentals of how universal approximators generalize about their inputs and access state space modalities. Second, the formal undecidability result concerning computable learnability of state spaces is proved. Additionally, we briefly comment on alternative models of learnability and hint towards their limits as well and discuss implications of our undecidability result. We are motivated by finding a formal limit of scientific exploration supported by machine learning that will frame the growing number of empirical results. Before giving the main result, key differences between the two discussed accesses to state space modalities are recapped.

2. Fundamentals of counterfactual conditionals and universal approximation

Both counterfactual and universal approximation access to state space modalities share the generalization from observations as a way allowing them to account for possible states of the physical system. The possible states range from actual states with non-actual properties to predicted, yet so far unobserved, novel states. In order to produce new observations, counterfactual access to state space modalities requires interventions in antecedents of counterfactual conditionals. Interventions aim to discover invariances, constant properties of different observations, that determine which counterfactuals can be accounted for by a generalization from observations of a physical system (cf. [26, 57], pp. 235–239). Universal approximation access to state space modalities based on ANNs does not require counterfactuals.

The difference between counterfactual and universal approximation access to state space modalities lies in how scientists and machine learning models learn invariances from observations. When relying on counterfactuals, scientists intervene on counterfactuals' antecedents to cause changes in the state space and evaluate the results with respect to theories about the physical system. When relying on the universal approximation capability of ANNs, there are no interventions, just observations sampled from a probability distribution over the state space determining the likelihood of individual observations. The fixed distribution makes it possible to guarantee a degree of reliability of inferences on unobserved samples and is among necessary requirements for generalization about the state space. Controlled by some risk minimization technique, universal approximators like ANNs learn from observations sampled from the distribution. The aim is to train an ANN that minimizes the risk estimated as the generalization error on a set of unseen observations from the state space (e.g., Empirical Risk Minimization, [54]). Observations are sampled in an independent and identical way (i.i.d., ibid.) or could be exchangeable [3]. Otherwise, the ANN lacks one of the necessary requirements for generalizing from observations and with it also guarantees regarding its inductive risk on the state space. In practice, what is an in- and out-of-distribution sample might be blurred due to big training datasets. This often leads to a surprisingly good performance on presumably out-of-distribution samples. It is not uncommon that these samples were, in fact, at least partially present in the training data. In other words, to see a clear-cut realization of theoretical limits in practice is not easy. This does not mean that the theoretical limits no longer apply. They are necessary, otherwise, it would not be possible to prove anything about learnability.

Invariances learned from counterfactuals capture interventions conceived by scientists (cf. [26], p. 198). Invariances learned by universal approximators capture unobserved and possibly also unconceived (from the human point of view, see, e.g., [48]) states of the physical system provided that the distribution over observations is fixed and observations are sampled in an exchangeable or i.i.d. manner during training and use of the ANN. These two types of invariances can complement each other when used to access state space modalities. The difference between them derives from the causal/non-causal inference split. Scientists using both types of inferences are able to answer causal as well as non-causal questions about state space modalities. The combination represents a form of epistemic enhancement in the sense of computational science [28] relying on machine learning models.

3. Universal approximation access to state space modalities

In order to generalize from observations, ANNs like any other machine learning method aim to turn invariant attributes of observations into features that are useful for performing correct inferences on yet unobserved samples. Modern deep, overparametrized ANNs [10,35] with strong approximation capabilities [19] seem to avoid harmful (as opposed to benign) overfitting of observations [9]. Gradient-based optimization techniques often enable overparametrized ANNs to fit training data (observations) exactly. This situation corresponds to interpolating training data, which still allows overparametrized ANNs to accurately generalize (ibid.). In order to use the standard apparatus of the (computable) learning theory, we do not diverge from the standard framework of independent and identically distributed or exchangeable samples from a fixed distribution on the state space of the physical system.

Representations consisting of features learned by deep ANNs during training involve invariances that establish universal approximation access to state space modalities. Improved approximation and, therefore, representational capabilities of deep ANNs compared to shallow ANNs are epistemically significant. The difference between shallow (a single hidden layer) and deep ANNs (several hidden layers) lies in the fact that the former is ineffective in representing observations that require approximating a hierarchy of non-linear functions, a shortcoming addressed by deep ANNs ([35], pp. 437–438). Errors in function approximation lead to inferential errors that are measured as the generalization error, i.e., the error rate of an ANN on new observations with respect to the ground-truth (e.g., failures to predict the correct label of an image). A large generalization error means that invariances represented by features learned by the ANN during training do not provide access to the target state space and that inferences supported by that ANN fail to provide access to state space modalities. For deep ANNs, large generalization errors result from overfitting training data (observations) in a harmful way that makes into features data attributes that do not exist outside the training dataset, failing to provide reliable access to the state space. Modern deep ANNs provide the best available universal approximation access to state space modalities. The ANN-based access to a state space of some physical system follows from the ability of universal approximators to learn a probably approximately correct hypothesis h that minimizes the generalization error on the state space.

The generalization error (or gap) of a predictor (hypothesis) h corresponding to a trained ANN is defined as ([42], Def. 8):

$$\Delta_{gen-error}(h) = R[h] - \widehat{R}_S[h],$$

where $\widehat{R}_S[h]$ is the empirical risk of h defined as h's average error on a set of i.i.d. observations $S = \left((x_1, y_1), \ldots, (x_m, y_m)\right)$ sampled from a fixed distribution over the state space and $R[h]$ is h's risk defined as expectation of the h's error over new observations sampled from the distribution. The predictor h belongs to a family of functions $\mathscr{H}$ that can be computed by an ANN, depending on its architecture and size ([5], p. 94). Learning theories such as Valiant's [52] Probably Approximately Correct framework define conditions for learnability of $\mathscr{H}$. Learnability of a system's state space can be characterized by the fact that there is a set of the state space observations $C = \{x_1, \ldots, x_m\}$ such that for any function $f : C \rightarrow \{0, 1\}$ we can find a predictor $h \in \mathscr{H}$ whose predictions on all observation from C correspond to f's predictions, that is, $\forall x \in C$, $h(x) = f(x)$. The size of the largest set of state space observations for which this is true is the VC-dimension [53] of the family of functions $\mathscr{H}$. If we require the predictors $h \in \mathscr{H}$ to be computable, which is natural considering that $\mathscr{H}$ corresponds to functions that can be computed by an ANN, finiteness of $\mathrm{VCdim}(\mathscr{H})$ does not guarantee learnability of $\mathscr{H}$ in every situation. Even more profoundly, in certain circumstances

relevant for learnability of state spaces finiteness of VCdim($\mathscr{H}$) is undecidable. In what follows, we provide details on how to derive these two results.

Formally, access to the state space is justified by PAC (Probably Approximately Correct) learnability of a hypothesis class $\mathscr{H}$ ([44,52], Def. 3.1). The universal approximator (an ANN) seeks to learn a function $h \in \mathscr{H} : \mathscr{X} \rightarrow \mathscr{Y}$, which maps i.i.d. observations $\{x_1, \ldots, x_n\}$ from a fixed distribution $\mathscr{D}$ over the state space $\mathscr{X}$ of a physical system to a set of labels $\{y_1, \ldots, y_n\}$ from the label space $\mathscr{Y}$. We define the label space as $\mathscr{Y} = \{0, 1\}$. The distribution $\mathscr{D}$ is defined over a σ-algebra consisting of subsets of the state space $\mathscr{X}$. If the hypothesis class $\mathscr{H}$ is learnable, the hypothesis h implements a binary classifier able to correctly answer membership queries about observations with respect to the subsets of the state space $\mathscr{X}$. It does so by learning invariances from training observations that capture the state space modalities.

Given the accuracy parameter $\epsilon \in (0, 1)$ and the confidence parameter $\delta \in (0, 1)$, the hypothesis h is learned by a universal approximator with the probability

$$\Pr[L(h) \leq \epsilon] \geq 1 - \delta \tag{1}$$

where $L(h) = \Pr_{x \sim \mathscr{D}}[h(x) \neq f(x)]$ and $f : \mathscr{X} \rightarrow \{0, 1\}$ is the ground-truth function for observations x of the state space $\mathscr{X}$ (ibid.). Inequality (1) holds for an i.i.d. set of observations of at least the size of $m_{\mathscr{H}}(\epsilon, \delta) \rightarrow \mathbb{N}$, with m determining the sample complexity required to learn $\mathscr{H}$ and output a probably approximately correct h as close as possible to the optimal classifier h^* such that $\Pr_{x \sim \mathscr{D}}[h^*(x) = f(x)] = 1$ ([44], Def. 2.1). In case $h, h^* \in \mathscr{H}$ and $L_{(\mathscr{D}, f)}(h^*) = 0$, the access to state space modalities is realizable and proper compared to nonrealizable and improper setting where $h, h^* \notin \mathscr{H}$ and learnability becomes agnostic ([44], Def. 3.3) and representation independent ([44], Remark 3.2). Agnostic PAC learnability changes inequality (1) as follows ([44], Def. 3.4):

$$\Pr\left[L_{\mathscr{D}}(h) \leq \min_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') + \epsilon\right] \geq 1 - \delta \tag{2}$$

PAC learnability requires a characterization of sample complexity, allowing us to determine which hypothesis classes are learnable. For binary classifiers with the $0-1$ loss function, this characterization is provided by VC-dimension [53]. The VC-dimension of a hypothesis class $\mathscr{H}$, VCdim($\mathscr{H}$), is determined using the ability of $\mathscr{H}$ to shatter a finite set $C \subset \mathscr{X}$. Shattering is defined as follows: A hypothesis class $\mathscr{H}$ shatters C if $\mathscr{H}$ contains all functions $C \rightarrow \{0, 1\}$, i.e., $|\mathscr{H}_C| = 2^{|C|}$ ([44], Def. 6.3). VCdim($\mathscr{H}$) is the size of the largest set $C \subset \mathscr{X}$ shattered by $\mathscr{H}$ ([44], Def. 6.5). Infinite VCdim($\mathscr{H}$) allows $\mathscr{H}$ to shatter arbitrarily large C and causes $\mathscr{H}$ not to be PAC learnable ([44], Theorem 6.6). Using Shalev-Shwartz and Ben-David's [44] proof of their Theorem 6.6 and following the logic of their Corollary 6.4 (ibid.), infinite VCdim($\mathscr{H}$) prevents learning $\mathscr{H}$ using a set of m observations because $\mathscr{H}$ also shatters a set of $2m$ observations, where the smaller set m does not provide ground-truths for observations from the larger set $2m$. As a result, the outputted classifier h is not a probably approximately correct hypothesis about the state space because it cannot correctly predict whether a sample from the larger set $2m$ belongs a finite set $C \subset \mathscr{X}$ and is, as a result, an observation of the state space $\mathscr{X}$. This means that a universal approximator with infinite VC-dimension lacks the epistemic justification to access state space modalities. Only finite VC-dimension can epistemically justify universal approximation access to the state space of a physical system, as finiteness of VC-dimension of hypothesis classes guarantees their PAC learnability according to Inequality (1).

Formally, No Free Lunch Theorem (NFL, [44], Theorem 5.1 and its computable version by [1], Lemma 19) can be used to show that some hypothesis classes are not PAC learnable and, as a result, algorithms based on universal approximation cannot access state space modalities. Let $\mathscr{X}$ be the state space of a physical system and $\mathscr{H}_{state-space}$ a hypothesis class containing computable functions $h : \mathscr{X} \rightarrow \{0, 1\}$. For any distribution $\mathscr{D}$ over a σ-algebra of subsets of the state space $\mathscr{X}$, an i.i.d. sample $S \sim \mathscr{D}^m$ of m state space observations, and $m = m_{\mathscr{H}}(\epsilon, \delta)$, where $\epsilon, \delta \in (0, 1)$, there is a computable algorithm $\mathscr{A}$ with $\Pr_{S \sim \mathscr{D}^m}\left[L_{\mathscr{D}}(\mathscr{A}(S)) \leq \min_{h \in \mathscr{H}_{state-space}}(L_{\mathscr{D}}(h) + \epsilon)\right] \geq 1 - \delta$. At the same time, there is also a subset of the state space $X = \{x_1, \ldots, x_{2m}\} \subseteq \mathscr{X}$ containing $2m$ observations and a computable function $g : \{x_1, \ldots, x_{2m}\} \rightarrow \{0, 1\}$ producing a uniform distribution $\widehat{\mathscr{D}}$ over $\{(x_1, g(x_1)), \ldots, (x_{2m}, g(x_{2m}))\}$. Setting $\epsilon < 1/8$ and $\delta < 1/7$, if the learner $\mathscr{A}$ is given an i.i.d. sample $S \sim \widehat{\mathscr{D}}^m$ of m state space observations, it fails with probability $\Pr_{S \sim \widehat{\mathscr{D}}^m}[L_{\widehat{\mathscr{D}}}(\mathscr{A}(S)) > 1/8] > 1/7$. This follows from the fact that $|\mathscr{X}| > 2m$ ([44], Corollary 5.2) and $\widehat{\mathscr{D}}$ exists because it is possible to computably find g ([1], Lemma 19). As a result, the hypothesis class $\mathscr{H}_{state-space}$ is not PAC learnable by the algorithm $\mathscr{A}$. Above, we already introduced some computability requirements of PAC learnability. We will now proceed with their full explanation.

3.1. Computability of universal approximation access to state space modalities

We provided the first set of conditions that allow a universal approximator with limited access (defined by sample complexity) to observations of the state space of a physical system to correctly predict whether new observations belong to some subset $X \subseteq \mathscr{X}$ of the state space or not. This characterization is based on statistical learning. The second set of conditions determines computability of universal approximation access to state space modalities. Traditionally, it is required that the runtime of the algorithm $\mathscr{A}$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ and in the size of representation of the hypotheses in $\mathscr{H}$ ([52], p. 1136; [33]). Recently, Agarwal et al. [1] initiated a study of computable PAC (CPAC) learnability that does not require learners $\mathscr{A}$ with polynomial runtime but learners that are just computable functions.

Formally, CPAC requires a computable algorithm $\mathscr{A}$ to output hypotheses (predictors) $h \in \mathscr{H}$ such that they are computable functions that can be evaluated on each input from the state space $\mathscr{X}$ ([1], Def. 8). Both realizable and agnostic as well as proper and improper learning settings are considered ([1], p. 5). Predictors $h \in \mathscr{H}$ produced by the computable learner $\mathscr{A}$ infer whether an observation $x \sim \mathscr{D}$ belongs to some subset $X \subseteq \mathscr{X}$ of the state space. Agarwal et al. [1] showed that PAC learnability of a hypothesis class $\mathscr{H}$ implies its CPAC learnability in both realizable proper and improper settings in case $\mathscr{H}$ has finite VC-dimension and is decidably representable (DR, ibid.). A hypothesis class $\mathscr{H}$ is DR if the set of all functions computed by an algorithm $\mathscr{A}$, which is from a set of decidable programs $\mathscr{P}$, equals $\mathscr{H}$ ([1], Def. 5). DR implies a weaker property of hypothesis classes, recursively enumerable representation (RER), which is implied by DR by $\mathscr{P}$ being a recursively enumerable set of programs ([1], Def. 6).

For a RER $\mathscr{H}$ with finite VC-dimension, $\mathscr{P}$ contains an algorithm $\mathscr{A}$ corresponding to the ERM (Empirical Risk Minimizer) that will evaluate each $h \in \mathscr{H}$ on an i.i.d. sample $S \sim \mathscr{D}^m$ of at least the size $m_{\mathscr{H}}(\epsilon, \delta)$ until it finds h^* satisfying Inequality (1) by $L_{(\mathscr{D}, f)}(h^*) = 0$, as we are in the realizable setting ([1], Proof of Theorem 10). In this case, $\mathscr{A}$ CPAC-learned the hypothesis class $\mathscr{H}_{state-space}$ by outputting h^*, a computable classifier that has access to state space modalities. If the hypothesis h^* is approximated by a neural network, the network's access to the state space modalities of a physical system follows from its universal approximation capability and the fact that (1) $\mathscr{H}_{state-space}$ is RER (or DR), (2) the distribution $\mathscr{D}$ over the σ-algebra of subsets of the state space $\mathscr{X}$ is fixed, (3) the sample $S \sim \mathscr{D}^m$ is i.i.d., and (4) h^* achieves $L_{(\mathscr{D}, f)}(h^*) = 0$. This is not the case in the agnostic proper setting (Inequality (2) with $h, h' \in \mathscr{H}$ and no h' s.t. $L_{(\mathscr{D}, f)}(h') = 0$) even though the hypothesis class $\mathscr{H}_{state-space}$ has finite VC-dimension and is decidably representable ([1], Theorem 11).

Definition 1. To show this, let us define a state space $\mathscr{X}_b$ containing b possible observations of the physical system. Then, there is a DR

hypothesis class $\mathscr{H}_{\mathscr{X}_b}$ with finite VC-dimension $\mathrm{VCdim}(\mathscr{H}_{\mathscr{X}_b}) \leq b$ that contains all hypotheses predicting at most b observations as belonging to some subset $X \subseteq \mathscr{X}$ of the state space.

Using the class $\mathscr{H}_{\mathscr{X}_b}$ from Definition 1, a proper CPAC learner $\mathscr{A}$ that can only learn subclasses $\mathscr{H}_b \subset \mathscr{H}_{\mathscr{X}_b}$ and, thus, can only output $h \in \mathscr{H}_b$ cannot learn modalities of the state space $\mathscr{X}_b$ in the agnostic setting. This is because no $h \in \mathscr{H}_b$ achieves $L_{(\mathscr{D},f)}(h) = 0$.

Corollary 1. Hypotheses from subclasses $\mathscr{H}_b \subset \mathscr{H}_{\mathscr{X}_b}$ cannot have full access to the state space in agnostic settings because its learnability is improper with respect to the subclasses, i.e., a hypothesis h s.t. $L_{(\mathscr{D},f)}(h) = 0$ does not belong to the subclass $h \notin \mathscr{H}_b$ but is contained in the state space hypothesis class $h \in \mathscr{H}_{\mathscr{X}_b}$. Even though the subclasses are DR and have finite VC-dimension, computable learnability of the state space $\mathscr{X}_b$ is not possible in the agnostic setting ([1], Theorem 11 for the original impossibility result and [50], pp. 6–7 for additional context). If a neural network learns $h \in \mathscr{H}_b$ in the agnostic setting, it does not provide universal approximation access to the state space modalities of $\mathscr{X}_b$. This is a result of the fact that proper PAC learnability of hypothesis classes does not imply computable proper learnability in the agnostic setting (ibid.).

This result strengthens the dependence of CPAC learnability of state space modalities on the realizability assumption about the hypothesis class $\mathscr{H}_{\mathscr{X}_b}$. That is, for both proper and improper setting, there is a hypothesis $h \in \mathscr{H}_b$ or $h \in \mathscr{H}_{\mathscr{X}_b}$ respectively such that $L_{(\mathscr{D},f)}(h) = 0$. In the realizable setting, any DR or RER hypothesis class with finite VC-dimension is CPAC learnable with a universal approximator $\mathscr{A}$ that implements an ERM and outputs the classifier based on h such that $L_{(\mathscr{D},f)}(h) = 0$. The classifier has access to state space modalities and can predict whether an observation x belongs to some subset $X \subseteq \mathscr{X}$ of the state space. There are several undecidability results regarding (C)PAC learnability that challenge this conclusion.

4. Undecidability of universal approximation access to state space modalities

Ben-David et al. [7,8] used the independence of the continuum hypothesis from the ZFC axioms to show that learnability treated as monotone compression remains undecidable. In case the continuum hypothesis is true, the cardinality of $[0,1]$ is $\aleph_1$ and, as a result, there are monotone compression schemes that can characterize learnability of $\mathscr{F}^*$, the family of all finite subsets of $[0,1]$ ([8], p. 47). If the continuum hypothesis is false, no such monotone compression scheme exists (ibid.). Since the continuum hypothesis is independent of ZFC, it cannot be proved nor refuted, and this makes learnability based on monotone compression undecidable. The independence of monotone compression from the ZFC set theory motivated the introduction of CPAC by Agarwal et al. [1], as replacing arbitrary functions with computable ones was considered to circumvent the independence of learnability from ZFC (provided ZFC is consistent).

Caro [15] showed that the key ingredient in different theories of learnability, a complexity measure such as VC-dimension, is susceptible to two types of undecidability. The first prevents proving learnability of computable hypothesis classes and the second prevents an algorithm to decide whether computable hypothesis classes are learnable (ibid.). In the former case, a recursively enumerable formal system $\mathscr{T}$ is called *consistent* if $\mathrm{VCdim}(\mathscr{H}_{\mathscr{T}}) < \infty$ ([15], Corollary 2.9), with each $h \in \mathscr{H}_{\mathscr{T}}$ representing a function that determines correspondence between the Gödel number of a theorem and the Gödel number of its negation ([15], Def. 2.4). Since in an *inconsistent* $\mathscr{T}$ the number of provable theorems is infinite, so is $\mathrm{VCdim}(\mathscr{H}_{\mathscr{T}}) = \infty$. By entailing finite VC-dimension, a recursively enumerable formal system $\mathscr{T}$ with $\mathrm{VCdim}(\mathscr{H}_{\mathscr{T}}) < \infty$ implies its own consistency. As a result, the finiteness of $\mathrm{VCdim}(\mathscr{H}_{\mathscr{T}})$ cannot be proved in $\mathscr{T}$, as this would mean that $\mathscr{T}$ can prove its own consistency, violating Gödel's second incompleteness theorem ([15], Corollary 2.11). Similarly, a Turing machine, deciding whether a computable hypothesis class has finite VC-dimension, does not exist, as this would mean that the halting problem is decidable ([15], Corollary 2.20). Finally, Sterkenburg ([50], Proposition 11) showed that Rice's Theorem can be used to derive undecidability of learnability of computable hypothesis classes as well by establishing correspondence between incomputable nontrivial index sets, i.e., $I \neq \emptyset$ or $I \neq \mathbb{N}$, and maximal computable families of hypothesis classes, i.e., $\mathrm{H} = \{\mathscr{H}_i\}_{i \in \mathbb{N}}$.

Apart from undecidability of VC-dimension finiteness, it was also shown that VC-dimension cannot be computed [22,39] nor approximately computed in polynomial time [37,38], assuming a version of the Exponential Time Hypothesis.

Corollary 2. Although there is a computable algorithm $\mathscr{A}$ that in the realizable setting outputs a classifier h computable on all observations from $\mathscr{X}$ such that $L_{(\mathscr{D},f)}(h) = 0$, neither PAC nor CPAC characterizes universal approximation access to state space modalities because finiteness of VC-dimension of hypothesis classes is undecidable and VC-dimension cannot even be approximated in polynomial time. Undecidability of VC-dimension finiteness can be expressed both as undecidability of learnability and independence of learnability from an axiom system underlying the state space $\mathscr{X}$.

We are now ready to give our main theorem about undecidability of universal approximation access to state space modalities.

Theorem 1. Let X and Z be sets such that $X = \{x_1, \ldots, x_n\}, X \subseteq \mathscr{X}$, and $Z = \{z_1, \ldots, z_n\}$ and $n \in \mathbb{N}$. Further, let $h \in \mathscr{H}_{\mathscr{X}}$ be a binary classifier implemented in an ANN such that $L_{(\mathscr{D},f)}(h) = 0$ that can correctly predict whether an observation z_i belongs to a subset X of the state space $\mathscr{X}$. By ZFC Axiom Schema of Separation, we have that $\forall Z \exists X (z_i \in X \leftrightarrow (z_i \in Z \wedge h(z_i)))$ All observations z for which h returns 1 form a set of observations X which is a subset of the state space $X \subseteq \mathscr{X}$. This relation is, however, undecidable.

Proof. Provided that ZFC is consistent and its axioms used for a basic characterization of the state space $\mathscr{X}$ of a physical system, the proof is constructed as follows. Universal approximation access to the state space $\mathscr{X}$ depends on a hypothesis $h \in \mathscr{H}_{\mathscr{X}}$ such that $L_{(\mathscr{D},f)}(h) = 0$. For a RER hypothesis class $\mathscr{H}_{\mathscr{X}}$ to be CPAC learnable by a universal approximator $\mathscr{A}$ with $\Pr_{S \sim \mathscr{D}^m}[L_{\mathscr{D}}(\mathscr{A}(S)) \leq L_{\mathscr{D}}(h) + \epsilon] \geq 1 - \delta$, $\mathscr{H}_{\mathscr{X}}$'s VC-dimension has to be finite $\mathrm{VCdim}(\mathscr{H}_{\mathscr{X}}) < \infty$ and $\mathscr{A}$ implement an ERM. Each observation z corresponds to a theorem p_z with Gödel number $\varphi(E_1^2(p_z))$ derived in a formal system[1] $\mathscr{T}$, where E^2 is a recursive enumeration of $\mathbb{N}^2$ given as a total bijective function $E_i^2 : \mathbb{N} \to \mathbb{N}, i = 1, 2$ ([15], pp. 7–8). If $\varphi(E_1^2(p_z)) = \neg\varphi(E_2^2(p_z))$, the formal system $\mathscr{T}$ underlying the state space $\mathscr{X}$ is inconsistent ([15], Def. 2.4) because Gödel number of the theorem p_z and of its negation coincide. VC-dimension of the hypothesis class $\mathscr{T}_{\mathscr{X}}$ of the inconsistent formal system $\mathscr{T}$ (see [15] for the definition of a formal system's hypothesis class) is infinite because from a contradiction, anything can be derived ([15], Theorem 2.7). By requiring that $\mathrm{VCdim}(\mathscr{H}_{\mathscr{X}}) < \infty$, we also require that the formal system $\mathscr{T}$ is consistent. Otherwise, there would be an explosion of derivable theorems about the state space $\mathscr{X}$ caused by contradictions in an inconsistent $\mathscr{T}$. When we attempt to use the formal system $\mathscr{T}$ to derive that $\mathrm{VCdim}(\mathscr{T}_{\mathscr{X}}) < \infty \Rightarrow \mathrm{VCdim}(\mathscr{H}_{\mathscr{X}}) < \infty$, Gödel's second incompleteness theorem is violated and $\mathscr{T}$ made inconsistent by proving its own consistency. Because we know that VC-dimension of an inconsistent $\mathscr{T}$ is infinite, we can derive that $\mathrm{VCdim}(\mathscr{T}_{\mathscr{X}}) = \infty \Rightarrow \mathrm{VCdim}(\mathscr{H}_{\mathscr{X}}) = \infty$. As a result, $\mathscr{H}_{\mathscr{X}}$ is not CPAC learnable by the universal approximator $\mathscr{A}$. This implies that universal approximation access to modalities of the state space $\mathscr{X}$ of a physical system is undecidable because the proof of $\mathrm{VCdim}(\mathscr{T}_{\mathscr{X}}) < \infty \Rightarrow \mathrm{VCdim}(\mathscr{H}_{\mathscr{X}}) < \infty$ allowing CPAC learnability of $\mathscr{H}_{\mathscr{X}}$ cannot be obtained in $\mathscr{T}$, that is, in

[1] $\mathscr{T}$ is recursively enumerable and supports arithmetic that allows it to derive infinitely many theorems.

the formal system describing the state space $\mathscr{X}$. ∎

Hanneke and Yang ([24], Section 1.7) recently provided a helpful perspective on this type of undecidability. When turning potential observations of the state space $\mathscr{X}$ into theorems of the formal system $\mathscr{T}$, for each observation z and corresponding theorem p_z, the hypothesis class $\mathscr{T}_{\mathscr{X}}$ returns a function $f_\theta(p_z)$, where θ is a parametrization of the function. As noted by Hanneke and Yang [24], each couple of (θ, p_z) represents a query against the hypothesis class $\mathscr{T}_{\mathscr{X}}$ which is opaque. The opaqueness is caused by the fact that the query access cannot be used to determine whether the class collapses to a single function $f_\theta(p_z) = 0$ whose indexes correspond to all possible parameter values or is infinite (ibid.) due to the contradiction $\varphi(E_1^2(p_z)) = \neg\varphi(E_2^2(p_z))$.

Corollary 3. Formally, a version of Agarwal et al.'s (2020, Definition 21) distinguishability problem fails with respect to the hypothesis class $\mathscr{H}_{\mathscr{T}_{\mathscr{X}}}$. A hypothesis $h \in \mathscr{H}_{\mathscr{T}_{\mathscr{X}}}$ takes a query $q = (\theta, p_z) \sim \mathscr{D}$ from a distribution over the query domain $\mathscr{Q}$ and outputs a function $f_\theta \in \mathscr{T}_{\mathscr{X}}$. There is no computable learner for $\mathscr{H}_{\mathscr{T}_{\mathscr{X}}}$ $\mathscr{A} : \bigcup_{n\in\mathbb{N}}(\mathscr{Q} \times f_\theta \in \mathscr{T}_{\mathscr{X}})^n \rightarrow \{single, infinite\}$ that is for any $\delta > 0$ and $S = (q_1, \ldots, q_m) \in \mathscr{Q}^m$, $m \in \mathbb{N}$, capable of predicting with probability at least $1 - \delta$ that $\mathscr{H}_{\mathscr{T}_{\mathscr{X}}}$ collapses to a redundantly indexed single function, $f_\theta(p_z) = 0$, or is infinite, i.e., each contradiction $\varphi(E_1^2(p_z)) = \neg\varphi(E_2^2(p_z))$ produces a new function based on $\theta = (\theta(k))_{k\in\mathbb{N}} \in c_c(\{0,1\})$. The equality $\varphi(E_1^2(p_z)) = \neg\varphi(E_2^2(p_z))$ can be satisfied infinitely many times. This makes $\mathrm{VCdim}(\mathscr{H}_{\mathscr{T}_{\mathscr{X}}}) = \infty$ and the distinguishability problem not CPAC learnable.

4.1. Large VC-dimension and other learning models

Deep ANNs are overparametrized and have enough capacity to memorize all training samples [58]. VC-dimension of hypothesis classes learnable by ANNs capable of memorization grows absurdly large and the memorization should prevent ANNs from generalizing about state spaces. Is there a difference between a hypothesis class with absurdly large VC-dimension and a hypothesis class with infinite VC-dimension? In the absurdly large case, the received wisdom is that an ANN learning such a class will be unable to generalize. In the infinite case, the class is not learnable which means that it is a priori clear that the ANN will not generalize either.

The study of interpolation, a situation in which an overparametrized ANN fits even noisy data perfectly and is still able to generalize [9], shows that the received wisdom is not applicable to deep ANNs. The key for understanding overparameterization seems to be that instead of a single or no hypothesis $h \in \mathscr{H}$ achieving zero empirical risk defined as $\mathscr{R}_{emp}(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i)$, there is now a set $\mathscr{S} = \{h \in \mathscr{H} : \mathscr{R}_{emp}(h) = 0\}$ of hypotheses that minimize the empirical risk equally well ([9], Sect. 3.6). Not all hypotheses from $\mathscr{S}$, however, generalize equally well. The learner $\mathscr{A}$, a combination of a neural architecture and an optimizer, capable of CPAC learning $\mathscr{H}$ approximates a hypothesis $h \in \mathscr{S}$ based on some inductive bias such as functional smoothness that helps the ANN generalize (ibid.). Although absurdly large VC-dimension of a hypothesis class does not play the traditional capacity control role that presumably helps the ANN to generalize, it is still finite and, thus, learnable in the interpolation regime. As a results, the difference between absurdly large and infinite VC-dimension of hypothesis classes is important, as only the latter prevents CPAC learnability.

We showed that the undecidability of learnability of state spaces is caused by a formal undecidability of finiteness of VC-dimension of hypothesis classes as well as by the fact that VC-dimension cannot be computed nor approximately computed in polynomial time, assuming a version of the Exponential Time Hypothesis. It is interesting to ask whether similar results hold for learnability in frameworks that do not use VC-dimension to measure the complexity of hypothesis classes. As far as learning frameworks go, (C)PAC, however, seems most natural because it describes batch learning that corresponds to many scientific machine learning workloads using ANNs.

An alternative is online learning, during which the learner predicts labels of sequentially arriving observations. In the realizable setting, there is a hypothesis $h^\star \in \mathscr{H} : \mathscr{X} \rightarrow \mathscr{Y}$ that generates the sequence of labels for the observations. Let $\mathscr{X}$ be the state space, $\mathscr{Y} = \{0,1\}$ the label space, and $\mathscr{A}$ an online learner. A sample $S = \{(x_t, y_t)\}_{t=1}^{T}$ represents a single run of the learner that at each time step $t \in [T \in \mathbb{N}]$ predicted $p_t = \mathscr{A}(S_{t-1}, x_t)$ and, as a result, made a number of mistakes given as $M_{\mathscr{A}}(S) = \sum_{t=1}^{T}\mathbb{I}_{[\mathscr{A}(S_{t-1}, x_t) \neq y_t]}$ [25]. Let us also define Littlestone Dimension (L-dimension, [36]) of a non-empty hypothesis class $\mathscr{H}$, $\mathrm{Ldim}(\mathscr{H})$, as the maximal $T \in \mathbb{N}$ that corresponds to the depth of a binary tree shattered by $\mathscr{H}$ ([6], Definition 8). A tree is shattered by $\mathscr{H}$ if for any root-to-leaf path $(x_1, y_1), \ldots, (x_T, y_T)$ a hypothesis $h \in \mathscr{H}$ exists such that $h(x_i) = y_i \; \forall i \; i \leq T$ ([6], Definition 7). If $\mathrm{Ldim}(\mathscr{H}) < \infty$, the hypothesis class $\mathscr{H}$ is online learnable. Further, if the learner $\mathscr{A}$ is Standard Optimal Algorithm (SOA, [6], Algorithm 1), $M_{SOA}(\mathscr{H}) = \mathrm{Ldim}(\mathscr{H})$, and there is no learner that can make less mistakes than SOA, that is $M_A(\mathscr{H}) \geq \mathrm{Ldim}(\mathscr{H})$ (ibid.). In the realizable setting, $\mathrm{Ldim}(\mathscr{H}_{\mathscr{X}}) < \infty$ characterizes online learnability of modalities of the state space $\mathscr{X}$ of some physical system (cf. ibid.).

This last statement can be contested using the same procedure that we applied to derive undecidability of learnability of state spaces under the PAC framework. First, Caro ([15], Proposition 4.5 and Corollary 4.6) showed 'Gödel' undecidability of finiteness of L-dimension similarly as for VC-dimension. Therefore, undecidability that holds for PAC learnability of state space modalities holds also for their online learnability. Second, as VC-dimension, L-dimension cannot be computed [22,39] nor approximately computed in polynomial time [37,38], assuming a version of the Exponential Time Hypothesis, which leads us to considering computable online learning. By constructing a hypothesis class with finite L-dimension for which no optimal online learner is computable, Hasrati and Ben-David ([25], Theorem 29) showed that L-dimension does not characterize computable online learning. We leave a detailed analysis of this result open. As mentioned, we focused on the PAC framework for its natural closeness to scientific machine learning workloads using ANNs.

4.2. Implications of the undecidability

We may now ask about the implications of the derived formal impossibility results for the ANN's use in everyday computational science. If hypotheses about the state space of a system are required to be functions computable by some ANN, finiteness of VC-dimension of the class of hypotheses expressible by the ANN does not guarantee learnability of the class in every situation (Corollary 1). An empirical consequence of this formal limit is that it is impossible to a priori prove that the trained ANN will correctly generalize about the state space in the agnostic CPAC setting, i.e., that it will reliably (within the bounds of the PAC definition of learnability) predict whether an observation belongs to the state space or not. This partial limit in the agnostic setting is, unfortunately, not the end of troubles because we showed that Caro's [15] method of proving undecidability of finiteness of VC dimension of hypothesis classes can be adapted to state space hypothesis classes (Theorem 1). This in turn means that finiteness of VC dimension of hypothesis classes cannot be used for characterizing learnability in any situation if we allow the connection between observations and theorems necessary for the proof of Theorem 1 to work as intended.

In a nutshell, using the popular and foundational (C)PAC framework, we cannot prove that ANNs will provide reliable access to the state space of a physical system. This is in contrast with the successful, everyday use of ANNs in computational science. One takeaway is that despite the fact we are unable to a priori prove that ANNs will generalize about the state space, they often do in a way satisfying the required level of robustness

and accuracy. This shows that the proof that cannot be derived in the (C) PAC framework is not about ANNs' functional properties as universal approximators but rather about justification of their generalization capability if we cannot rely on VC dimension.

Spelda and Stritecky [49] obtained an epistemic justification of the generalization capability of overparameterized ANNs by identifying conditions that allow a computable learner $\mathscr{A}$ to become locally stable. This kind of local stability is an empirical phenomenon, which depends on a set of conditions. The occurrence of local stability cannot be a priori proved and in the current setting is relevant only for large VC dimension caused by overparameterization, which prevents generalization error bounds independent of the data distribution from justifying the ANN's generalization capability. PAC learnability of hypothesis classes is distribution independent, and this led Bousquet et al. [11] to propose an alternative theory for the realizable setting that allows the rate of generalization error convergence to depend on the distribution over observations. The PAC framework, on the other hand, features a single, worst-case rate that applies uniformly to all distributions consistent the with hypothesis class (ibid.). In Bousquet et al.'s (2021, Theorem 1.9) theory of universal learning, the rate of convergence is controlled by finiteness of Littlestone and VC-Littlestone trees (VC-Littlestone tree combines the structures behind VC and L-dimension, [11], Def. 1.8) for the hypothesis class $\mathscr{H}$. A VC-Littlestone tree for $\mathscr{H}$ can be infinite, but then learnability of $\mathscr{H}$ requires arbitrarily slow rates of convergence of the generalization error, which means that $\mathscr{H}$ is learnable but the learner $\mathscr{A}$ cannot predict the rate of convergence ([11], p. 535).

It seems that if the state space of a physical system is characterized by a hypothesis class, theories defining learnability of hypothesis classes have problems predicting whether ANNs trained on observations of that state space will access it reliably. Judged by the success of machine learning in empirical science, this might be a manifestation of the gap that is currently separating learning theory from machine learning practice [20,23,31]. There is, however, an important distinction to be made. Our undecidability result challenges means that can be used to a priori justify reliable, universal approximation access to state spaces not the ability of universal approximators to learn from state space observations. Therefore, undecidability of learnability of state spaces makes it hard to predict whether ANNs will correctly generalize about state spaces of physical systems but does not prevent imperfect generalization if the empirical conditions are right. The problem is that the 'right' empirical conditions do not seem to be *fully* characterized by the requirements of (C)PAC learnability and due to this we cannot know how imperfect the generalization capability will be until all task-relevant observations have been seen. This means that it is hard to provide an epistemic justification for machine learning tools in scientific contexts, which, however, continue to work well thanks to careful experimental practices. Our undecidability result has a constructive side that further motivates philosophy of science questions asking how we justify the role of machine learning in science.

5. Conclusion

We showed that computable PAC learnability of hypothesis classes does not characterize access to state space modalities. Our result about undecidability means that modal knowledge about state spaces obtained using ANNs does not have formal guarantees. Undecidability of learnability of state spaces defines an epistemic limit of scientific inquiry aided by machine learning. The next step in fully understanding the role of machine learning in various scientific fields is to find out what are the ways in which formal limits can meaningfully engage with empirical exploration.

Funding

This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595).

CRediT authorship contribution statement

Petr Spelda: Conceptualization, Formal analysis, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing. **Vit Stritecky:** Investigation, Writing - Original Draft, Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595). We are thankful to the reviewers for the journal for their feedback that helped us to improve the paper.

References

[1] S. Agarwal, N. Ananthakrishnan, S. Ben-David, T. Lechner, R. Urner, On Learnability with Computable Learners, 31st Int. Conf. Algorithm Learn. Theory (2020).
[3] Arjovsky M., Bottou L., Gulrajani I., Lopez-Paz D. (2019) Invariant Risk Minimization. ⟨https://arxiv.org/abs/1907.02893⟩.
[4] A. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (3) (1993) 930–945.
[5] P.L. Bartlett, A. Montanari, A. Rakhlin, Deep learning: a statistical viewpoint, Acta Numer. 30 (2021) 87–201.
[6] S. Ben-David, D. Pál, S. Shalev-Shwartz, Agnostic online learning, Proc. 22nd Annu. Conf. Learn. Theory (2009).
[7] Ben-David S., Hrubeš P., Moran S., Shpilka A., Yehudayoff A. (2017) On a learning problem that is independent of the set theory ZFC axioms. ⟨https://arxiv.org/abs/1711.05195⟩.
[8] S. Ben-David, P. Hrubeš, S. Moran, A. Shpilka, A. Yehudayoff, Learnability can be undecidable, Nat. Mach. Intell. Vol. 1 (2019) 44–48.
[9] M. Belkin, Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation, Acta Numer. 30 (2021) 203–248.
[10] Y. Bengio, Y. LeCun, G. Hinton, Deep learning for AI, Commun. ACM 64 (7) (2021) 58–65.
[11] Bousquet O., Hanneke S., Moran S., van Handel R., Yehudayoff A. (2021) A theory of universal learning. In *STOC 2021: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*.
[12] Butter A., Plehn T., Schumann et al. (2022) Machine Learning and LHC Event Generation. ⟨https://arxiv.org/abs/2203.07460⟩.
[13] P. Calafiura, D. Rousseau, K. Terao, Artificial Intelligence for High Energy Physics, World Scientific Publishing Company, Singapore, 2022.
[14] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. 91 (4) (2019) 045002.
[15] Caro M. (2021) Undecidability of Learnability. ⟨https://arxiv.org/abs/2106.01382v2⟩.
[16] E. Cuoco, J. Powell, M. Cavaglià, K. Ackley, M. Bejger, C. Chatterjee, M. Coughlin, S. Coughlin, P. Easter, R. Essick, Enhancing gravitational-wave science with machine learning, Mach. Learn.: Sci. Technol. 2 (2020) 011002.
[17] CMS Collaboration, A deep neural network to search for new long-lived particles decaying to jets, Mach. Learn. Sci. Technol. 1 (2020) 035012.
[18] B.L. DeCost, J.R. Hattrick-Simpers, Z. Trautt, A.G. Kusne, E. Campo, M.L. Green, Scientific AI in materials science: a path to a sustainable and scalable paradigm, Mach. Learn.: Sci. Technol. 1 (2020) 033001.
[19] R. DeVore, B. Hanin, G. Petrova, Neural network approximation, Acta Numer. 30 (2021) 327–444.
[20] G.K. Dziugaite, A. Drouin, B. Neal, N. Rajkumar, E. Caballero, L. Wang, I. Mitliagkas, D.M. Roy, In search of robust measures of generalization, Adv. Neural Inf. Process. Syst. 33 (NeurIPS 2020) (2020).
[21] N. Fu, L. Wei, Y. Song, Q. Li, R. Xin, S.S. Omee, R. Dong, E.M.D. Siriwardane, J. Hu, Material transformers: deep learning language models for generative materials design, Mach. Learn.: Sci. Technol. 4 (2023) 015001.

[22] M. Frances, A. Litman, Optimal mistake bound learning is hard, Inf. Comput. 144 (1998) 66–82.
[23] Gastpar M., Nachum I., Shafer J., Weinberger T. (2024) Fantastic Generalization Measures are Nowhere to be Found. In *The Twelfth International Conference on Learning Representations*.
[24] S. Hanneke, L. Yang, Bandit Learnability can be Undecidable, Proc. 36th Annu. Conf. Learn. Theory PMLR 195 (2023) 1–38.
[25] N. Hasrati, S. Ben-David, On computable online learning, Proc. 34th Int. Conf. Algorithm Learn. Theory PMLR 201 (2023) 1–19.
[26] C. Hitchcock, J. Woodward, Explanatory generalizations, Part II: plumbing explanatory depths, NoûS. 37 (2) (2003) 181–199.
[27] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (5) (1989) 359–366.
[28] P. Humphreys, Extending Ourselves: Computational Science, Empiricism, and Scientific Method. New York, NY, Oxford University Press, 2004.
[29] A. Iess, E. Cuoco, F. Morawski, J. Powell, Core-collapse supernova gravitational-wave search and deep learning classification, Mach. Learn. Sci. Technol. 1 (2020) 025014.
[30] B. Jalali, Y. Zhou, A. Kadambi, V. Roychowdhury, Physics-AI symbiosis, Mach. Learn.: Sci. Technol. 3 (2022) 041001.
[31] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, S. Bengio, Fantastic generalization measures and where to find them, Eighth Int. Conf. Learn. Represent. (2020).
[32] G. Kasieczka, T. Plehn, A. Butter, The Machine Learning landscape of top taggers, SciPost Phys. 7 (1) (2019) 014.
[33] M.J. Kearns, U.V. Vazirani, Cambridge, MA. An Introduction to Computational Learning Theory, The MIT Press, 1994.
[34] R. Koskinen, Kinds of modalities and modeling practices, Synthese 201 (2023) 196.
[35] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[36] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm, Mach. Learn. 2 (1988) 285–318.
[37] P. Manurangsi, A. Rubinstein, Inapproximability of VC Dimension and Littlestone's Dimension. Proc. 2017 Conf. Learn. Theory PMLR 65 (2017) 1432–1460.
[38] Manurangsi P. (2023) Improved Inapproximability of VC Dimension and Littlestone's Dimension via (Unbalanced) Biclique. In *The 14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*.
[39] C.H. Papadimitriou, M. Yannakakis, On limited nondeterminism and the complexity of the V-C Dimension, J. Comput. Syst. Sci. 53 (1996) 161–170.
[40] J. Pearl, The seven tools of causal inference, with reflections on machine learning, Commun. ACM 62 (3) (2018) 54–60.
[41] S. Ravanbakhsh, J. Oliva, S. Fromenteau, L.C. Price, S. Ho, J. Schneider, B. Póczos, Estimating Cosmological Parameters from the Dark Matter Distribution, Proc. Mach. Learn. Res. 48 (2016) 2407–2416.
[42] B. Recht, M. Hardt, Patterns, Predictions, and Actions: Foundations of Machine Learning, Princeton University Press, Princeton, NJ, 2022.
[43] Schäfer A.M., Zimmermann, H.G. (2006) Recurrent Neural Networks Are Universal Approximators. In S. Kollias et al. (Eds.), International Conference on Artificial Neural Networks (pp. 632-640). Berlin: Springer.
[44] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. New York, NY, Cambridge University Press, 2014.
[45] J. Shlomi, P. Battaglia, J.-R. Vlimant, Graph neural networks in particle physics, Mach. Learn.: Sci. Technol. 2 (2020) 021001.
[46] Y. Sjölin Wirling, T. Grüne-Yanoff, The epistemology of modal modelling, Philos. Compass 16 (10) (2021) e12775.
[47] Y. Sjölin Wirling, T. Grüne-Yanoff, Introduction to the Synthese topical collection 'Modal modeling in science: modal epistemology meets philosophy of science, Synthese 201 (2023) 208.
[48] Spelda P., Stritecky V. (2021) What Can Artificial Intelligence Do for Scientific Realism? *Axiomathes* 31, 85-104.
[49] P. Spelda, V. Stritecky, Why and how to construct an epistemic justification of machine learning? Synthese 204 (2024) 74.
[50] Sterkenburg T.F. (2022) On characterizations of learnability with computable learners. In *The 35th Annual Conference on Learning Theory*.
[51] T.E. Tahko, The modal basis of scientific modelling, Synthese 201 (2023) 75.
[52] L.G. Valiant, A Theory of the Learnable, Commun. ACM 27 (11) (1984) 1134–1142.
[53] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Its Appl. XVI (2) (1971) 264–280.
[54] V.N. Vapnik, Berlin. The Nature of Statistical Learning Theory, Springer, 1995.
[55] T. Williamson, Malden, MA: Blackwell Publishing, Philosophy Philosophy (2007).
[56] T. Williamson, Spaces of Possibility, R. Inst. Philos. Suppl. 82 (2018) 189–204.
[57] J. Woodward, Explanation and Invariance in the Special Sciences, Br. J. Philos. Sci. 51 (2) (1997) 197–254.
[58] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, Fifth Int. Conf. Learn. Represent. (ICLR) (2017).

Petr Spelda is an Assistant Professor at the Department of Security Studies, Charles University. His research focuses on safe machine learning and the building blocks of trustworthy artificial intelligence. He publishes in several fields, including computer and political science or philosophy. More about his interdisciplinary research can be found at https://research.spelda.cz/.

Vit Stritecky is an Associate Professor in International Security and Head of Department of Security Studies, Charles University. His research focuses on policy and regulatory issues connected to machine learning and on the role of standardization in addressing safety and security concerns about technology. His work ranges from strategic studies to sociotechnical understanding of artificial intelligence.