Human Induction in Machine Learning: A Survey of the Nexus

Preprint, forthcoming in *ACM Computing Surveys*[1]

Petr Spelda
Department of Security Studies, Faculty of Social Sciences, Charles University and prg.ai, Smetanovo nabrezi 6, Prague 1, 110 01, Czech Republic, petr.spelda@fsv.cuni.cz

Vit Stritecky
Department of Security Studies, Faculty of Social Sciences, Charles University, Smetanovo nabrezi 6, Prague 1, 110 01, Czech Republic, vit.stritecky@fsv.cuni.cz

Abstract

As our epistemic ambitions grow, the common and scientific endeavours are becoming increasingly dependent on Machine Learning (ML). The field rests on a single experimental paradigm, which consists of splitting the available data into a training and testing set and using the latter to measure how well the trained ML model generalises to unseen samples. If the model reaches acceptable accuracy, an a posteriori contract comes into effect between humans and the model, supposedly allowing its deployment to target environments. Yet the latter part of the contract depends on human inductive predictions or generalisations, which infer a uniformity between the trained ML model and the targets. The paper asks how we justify the contract between human and machine learning. It is argued that the justification becomes a pressing issue when we use ML to reach 'elsewheres' in space and time or deploy ML models in non-benign environments. The paper argues that the only viable version of the contract can be based on optimality (instead of on reliability which cannot be justified without circularity) and aligns this position with Schurz's optimality justification. It is shown that when dealing with inaccessible/unstable ground-truths ('elsewheres' and non-benign targets), the optimality justification undergoes a slight change, which should reflect critically on our epistemic ambitions. Therefore, the study of ML robustness should involve not only heuristics that lead to acceptable accuracies on testing sets. The justification of human inductive predictions or generalisations about the uniformity between ML models and targets should be included as well. Without it, the assumptions about inductive risk minimisation in ML are not addressed in full.

CCS CONCEPTS

General and reference – Surveys and overviews; Machine learning approaches; Machine learning theory – Inductive inference; Philosophical/theoretical foundations of artificial intelligence

Additional Keywords and Phrases

inductive reasoning, distribution shifts, robustness, empirical risk minimisation, invariant risk minimisation, epistemic justification of human-machine contracts, epistemology of machine learning

[1] http://dx.doi.org/10.1145/3444691.

1 Introduction

By delivering exceedingly powerful results, Machine Learning (ML), and its most successful paradigm presently available, Deep Learning (LeCun et al. 2015; Schmidhuber 2015), established itself as the focal point within the study of Artificial Intelligence (AI). The (seemingly) strong empirical results had emerged from the interplay between large volumes of available data and the continuous increases in computational resources, which reinvigorated the end-to-end generalisation learning by artificial neural networks. These rather recent developments continue to prompt all sorts of reactions. Some argue that they provide a renewed justification for the study of existential risks, which supposedly stem, among other things, from the *hypothetical* instances of AI such as Artificial General Intelligence (Sotala and Yampolskiy 2014) or from the largely disputed notion of Superintelligence (Bostrom 2014). Others are likely to feel that more practical concerns should dominate the study of consequences brought about by the exceedingly successful ML. Therefore, we observe a vibrant activity going on in AI ethics (Dignum 2018), in the study of transcending bias and achieving fair ML (Barocas et al. 2019), in explainable ML (Miller 2019), not to mention also the inquiries into safety and security of ML seeking to deliver robustness (Amodei et al. 2016). From a bird's-eye view, aware of the possibility that both the near- and far-term outcomes of ML/AI might be unfavourable (cf. Yampolskiy 2019), we seek a rapprochement guaranteeing the philosophical and technical alignment of ML/AI with human intentions and needs, securing our future flourishing.

Yet as we analysed the consequences, we lost track of varieties of assumptions in ML. They are hypotheses formed to justify that a ML model and some data would be together sufficient to achieve the desired aim. That is, learning generalisations which support inferences allowing the ML model to reliably perform the task at hand under various conditions. The assumptions reflect a variety of human epistemic viewpoints that inform how we reason about the plausibility/implausibility of machine learning in view of various empirical situations. Put differently, the assumptions are like a variety of hinges, each allowing for a different kind of learning, and thus serving as the nexus between anthropic and artificial. The hinges are where we put the inductive inferences and generalisations, which capture our expectations about the performance of ML models in different empirical situations.

The hinges, and thus the Human-Machine Nexus, are presently overshadowed by the plethora of earlier mentioned research areas. Therefore, the paper seeks to ameliorate this lack of attention and explain why it has become desirable to bring some light over to this domain. The lack of attention is, however, only partially the result of a scholarly neglect. First, the accessibility of the assumptions is made difficult by the fact that the epistemology of machine learning still receives comparatively less attention than other areas. There are only few works attempting to provide an epistemological treatment of statistical learning theory (cf. Harman and Kulkarni 2007; Corfield et al. 2009; Corfield 2010; von Luxburg and Schölkopf 2011; Spelda 2018), which, moreover, captures only a part of the story. Second, the recent history of ML has been dominated by an empiricist practice that derives estimates of performance from the a posteriori evaluation. More daunting problems connected to robustness of ML models under changing conditions (cf. e.g. Arjovsky et al. 2019) call this flavour of empiricism into question and provide a stage

for the renewed interest in varieties of assumptions in ML. Finally, the state-of-the-art deep artificial neural networks seems to defy classical results from statistical learning theory (cf. Zhang et al. 2017). The a priori guarantees (generalisation bounds) provided by the theory thus require further empirical qualifications (cf. Kawaguchi et al. 2019; Nagarajan and Kolter 2019). Though supposedly resolving the issue, the landscape of assumptions grows even more complex and becomes almost impenetrable to non-experts.

To advance understanding of varieties of human assumptions in ML, including their justification, the paper approaches the problem as a two-tier procedure. First, humans commit to assumptions, either inductive predictions or generalisations (first tier), which then enable ML models to learn generalisations (second tier) supposedly applicable beyond training data (the definitions of inductive prediction and generalisation are provided in Subsection 1.1). It is crucial to study the justification of the first tier human assumptions since they cause successes or failures of the trained ML models. Our contribution lies in showing what is the only viable option (Schurz's optimality justification of the first tier assumptions; the link between justification of inductive predictions/generalisations and Schurz's optimality, including its definition, is provided in Subsection 1.1) if we cannot justify that the human assumptions, inductive predictions or generalisations (first tier), can be considered reliable.

By showing that varieties of assumption in ML (first tier) can be rendered as human inductive predictions/generalisations, the paper seeks a clearer picture that would emphasise the human role in ML and elucidate the true shape of the Human-Machine Nexus. The rest of the paper is organised as follows. Subsection 1.1 offers a formal definition of the assumptions problem. Section 2 discusses the factors that make the exposition of varieties of assumptions in ML a timely endeavour. Section 3 then provides an analysis of the assumptions, which is directed towards the goal of Section 4. Section 4 outlines the distinction between the future reliability and present optimality of human inductive inferences and the consequence it has for ML in general and for ML in specific, epistemically challenging situations.

1.1 The Formal Definition of the Assumptions Problem

For the Definition 1.1 of inductive prediction and 1.2 of inductive generalisation, we rely directly on Gerhard Schurz's formulations 1 and 2 respectively (2019, p. 2):

> *1.1 Inductive prediction: r% of all so far observed Fs have been Gs. Therefore, with a (subjective) probability of approximately r%, the next F will be a G – and thus will be predicted to be a G, provided r is greater than 1/2 and F is the total evidence regarding the next observed individual.*
>
> *1.2 Inductive generalization: r% of all so far observed Fs have been Gs. Therefore, with high (subjective) probability, approximately r% of all Fs are Gs.*

Concerning the first tier (human) assumptions introduced above, `Gs` from Definition 1.1 and 1.2, whose probability `r` is of concern to us, represent target environments that preserved uniformity with training/evaluation environments used to produce the deployed ML model. A genuine justification of either of the inductive inferences on `Gs`,

performed by humans according to Definition 1.1 or 1.2, would give us a chance to estimate the likelihoods of distribution shifts, which disrupt uniformities between training/evaluation and target datasets.

Concerning the second tier (machine) inductive inferences, `Gs` from Definition 1.1 represent correct predictions produced by a trained ML model on individual samples from a holdout (test set) measured by some loss function and decision threshold. Once all predictions on the holdout samples have been collected, `r` from Definition 1.2 becomes an approximation of the ML model's generalisation performance according to the selected accuracy metric.

Proposition 1: For the second tier `r` from Definition 1.2 (of any trained ML model) to hold for targets past the holdout, the first tier `r`, either from Definition 1.1 or 1.2, must hold as well.

The first tier `r` reflects the level of support available to human predictions on (Definition 1.1) or generalisations about (Definition 1.2) distribution shifts past the observed targets. Even though at the first tier humans infer on targets while at the second ML models infer on samples from targets, the problem remains the same: the justification of projections from observed to unobserved. This has two immediate corollaries. First, if the concern is with the justification, the functional distinction between human and ML inductive predictions or generalisations ceases to play any role. Second, in light of Proposition 1, it needs to be determined whether a non-circular justification of the first tier `r` (for both Definition 1.1 and 1.2) is possible for *dynamic environments* which cause distribution shifts.

If statistical learning theory is used, which is the obvious starting point, the first tier `r` from Proposition 1 depends on probabilistic justifications which require samples to be either *independent and identically distributed* (Vapnik 1999) or *exchangeable* (i.e. ground-truths must be invariant to permutations of the samples which are, however, no longer required to be independent and identically distributed, cf. Arjovsky et al. 2019). The former requirement, which underpins Empirical Risk Minimisation, presupposes that training, holdout, and target samples come from a single, yet not necessarily known, probability distribution (Vapnik 1999; as per Definition 1.1, the uniformity among targets is predicted because the first tier `r`, based on samples from a single distribution, does not support generalisations about distribution shifts). The latter requirement, which underpins Invariant Risk Minimisation, presupposes a mixture of multiple distributions (Arjovsky et al. 2019; as per Definition 1.2, the uniformity among targets becomes a generalisation because the first tier `r`, based on samples from multiple distributions, presupposes a limit to distribution shifts). Training, holdout, and target samples are no longer required to come from a single distribution and to be independent and identically distributed (ibid.). However, to satisfy exchangeability ground-truths underlying the mixture must remain invariant to permutations of the samples. Gerhard Schurz (2019, Chapter 4) showed that these probabilistic justifications, underpinning Empirical and Invariant Risk Minimisation, are inductive inferences. Therefore, using statistical learning theory to justify the first tier `r` from Proposition 1 generates a circle and thus a failure for

our purposes. More profoundly, as noted by Vladimir Vovk (2015, pp. 163-64), exchangeability (as well as the 'independent and identically distributed' requirement) represents a ground-truth itself. Ergo, in case of inaccessible/unstable ground-truths ('elsewheres' and non-benign targets), which do not remain invariant under permutations of samples (nor produce independent and identically distributed samples), a non-circular justification of the first tier `r` from Proposition 1 becomes impossible.

However, the first tier `r` from Proposition 1 can be replaced with a meta-inductive evaluation of the first-order risk minimisation methods (candidate methods further) based on their past successes and failures recorded as uniformities and disuniformities among targets. Such a replacement will become independent of any ground-truth and can be acquired by proving that there is a long-run no regret risk minimisation strategy which can adapt to arbitrary distribution shifts. The strategy meeting such requirements is Schurz's exponential attractivity-weighted meta-induction (eMI, 2019). Its optimality is defined strictly in relation to the presently available success records. This allows eMI to avoid the inductive inferences involved in the first tier `r` from Proposition 1 (Definition 1.1 or 1.2) that cannot be justified under inaccessible/unstable ground-truths. Let us recap Schurz's result to provide the definition of optimality and its relation to online learning with expert advice.

Formally, Schurz's eMI satisfies the same (short-run) upper bound on regret (the worst case deterioration of eMI's predictive success rate on target uniformities compared to the maximal success rate of non-meta-inductive candidate methods) as Cesa-Bianchi's and Lugosi's exponentially weighted average forecaster (Schurz 2019, Theorem 6.9 (ii) and Cesa-Bianchi and Lugosi 2006, Theorem 2.3, pp. 17-20 respectively). Following Schurz (2019, Definition 6.4 for the class of weighted-average meta-induction; Section 6.6.2 for its version utilising exponential weights based on absolute successes of the candidate methods – 6.8 (i), also Schurz and Thorn 2020, Definition 9), for the round $n + 1$ eMI predicts a weighted average of the candidate methods' predictions:

$$\frac{\sum_{1\leq i\leq m} w_n(P_i) \cdot pred_{n+1}(P_i)}{\sum_{1\leq i\leq m} w_n(P_i)}$$

where `m` equals the number of candidate methods at the round `n`, $pred_{n+1}P_i$ corresponds to a candidate method prediction for the round $n + 1$, and $w_n(P_i)$ to the weight (the regret-based attractivity of P_i assigned by eMI) of that method in the round `n` defined as:

$$e^{\eta \cdot abs_n(P_i)}$$

where $abs_n(P_i)$ equals the absolute success of that method accumulated during past rounds until the round `n` (value between `[0, n]`). Setting the time-varying exponential potential

$$\eta = \sqrt{8 \cdot \ln(m)/(n+1)}$$

yields the upper bound on eMI's regret that holds uniformly over time for $n \geq 1$, assuming a loss function which receives values between `[0, 1]` and remains convex in its first

argument (Schurz 2019, Theorem 6.9 (ii); Schurz and Thorn 2020, Theorem 10 (i); Cesa-Bianchi and Lugosi 2006, Theorem 2.3):

$$\mathrm{maxsuc_n} - \mathrm{suc_n(eMI)} \leq 1.78 \cdot \sqrt{\ln(m)/n}$$

Since, considering the candidate methods, we are dealing with a maximisation of the success rate $\max_{1\leq i\leq m} \mathrm{abs_n}(P_i)/n$ rather than with a minimisation of the cumulative loss $\min_{1\leq i\leq m} L_n(P_i)$, as in Cesa-Bianchi's and Lugosi's (2006) Theorem 2.3, the upper bound on eMI's short-run regret remains equivalent to the upper bound given by Cesa-Bianchi's and Lugosi's (2006) Theorem 2.3 (also cf. Schurz 2008, p. 298). In the long run, eMI transforms into a no regret prediction strategy (Schurz 2019, Theorem 6.9 (iii)):

$$\lim_{n\to\infty} \sup(\mathrm{maxsuc_n} - \mathrm{suc_n(eMI)}) \leq 0$$

and becomes the *universally optimal* evaluation procedure of first-order risk minimisation methods free of any dependence on ground-truths and other presuppositions about uniformities among targets. Therefore, it can replace the first tier `r` from Proposition 1 and provide a solution to the problem of assumptions formally defined in this subsection and fully elaborated in the rest of the paper.

2 Elsewheres in Space & Time and Non-Benign Environments

Varieties of assumptions in ML become crucial if there is a likelihood that the training and evaluation of a ML model might incorrectly capture the phenomena which form the subject of the learning. In these situations, we rely on varieties of uniformity assumptions, which seek to predict that the next environments/datasets will be sufficiently like the data used to train/evaluate the ML model, thus providing an epistemic justification for its usage. According to the range afforded, these assumptions are either inductive predictions or generalisations (first tier). The former expects the ML model to be reliable in domains (immediately) adjacent to the training/evaluation data (first tier prediction), the latter presumes robustness across the whole range of domains that might contain the phenomena addressed by the learning in various contexts (first tier generalisation). In this sense, machine learning is as much a human endeavour as it is the effort to devise increasingly self-directed machine learning of generalisations. Therefore, the human part of the Nexus is equally, if not more, important as the progress of its artificial counterpart. This issue is most pressing in situations where the human uniformity assumptions remain insufficiently justified. By applying machine learning in these situations, ML models become exposed to distribution shifts between training/evaluation and target (deployment) data. The gap might diminish robustness of the ML models and produce inaccurate results. There are two particularly sensitive areas, which call for a clear-cut exposition of varieties of assumptions that constitute the Human-Machine Nexus. The first one is the application of ML models to reach elsewhere in space and time. In this case, we extend our reach by training ML models on noisy, incomplete or proxy data in order to enable empirical reconstructions of phenomena predicted by our theories (those phenomena are, however, directly unobservable). Recent examples of this practice can be

observed in astrophysics (Bouman et al. 2016; Fish et al. 2016) or geosciences (Haywood et al. 2009; Karpatne et al. 2018). The second area comprises deployments of ML models to non-benign environments populated by bounded and/or unbounded adversaries, who generate data aimed to maximise distribution shifts, thus diminishing robustness of the deployed ML models (Laskov and Lippmann 2010; Barreno et al. 2010; Amodei et al. 2016, pp. 16-20). The rest of this section intends to briefly introduce the most revealing applications of ML in these two areas allowing to build a connection to varieties of assumptions in ML, which will be analysed in Section 3.

2.1 Elsewheres in Space and Time

The first area is defined by using machine learning to reconstruct unobservable phenomena, which are distant from our present either in space or time. Such a practice seeks to extend reference frames of spatiotemporally situated observers by incorporating phenomena from remote epistemic locations. Generally, inductive inferences are based on inferring from the observed to the unobserved, regardless whether the unobserved lies in the past (i.e. retrodictions and generalisations [about the past]), in the future (i.e. predictions and generalisations [about the future]), or 'elsewheres' (inferences about spatially distant locations). The data that can be obtained from such remote epistemic locations are (extremely) sparse and noisy, in some cases representing mere proxies. As a result, the reconstructive tasks face empirical underdetermination. It produces an inverse and ill-posed learning problem described by a one-to-many relation between many possible phenomena fitting the available data. By training on some synthetic or natural data, both coming from the observer's present, the applied ML model seeks to learn an inductive schema capable of retrodiction in order to reconstruct the unobservable phenomenon. The training procedure seeks to establish prior assumptions restricting the set of possible phenomena that can be reconstructed from the available data (for visual data e.g. Zoran and Weiss 2011). The ML model reconstructs the phenomenon so that it is likely under the prior assumptions (ibid.). The assumptions are derived from synthetic or local data, in both cases reflecting our acquaintance with the *present* spatiotemporal location.

Whether we prefer to base the assumptions on synthetic data, stemming from the prevailing theoretical persuasion about the phenomenon, or natural data (e.g. natural scenes if reconstructing images) remains beside the point. Rather, the critical focus lies on a different kind of assumption. Namely, how we justify the inductive inference predicting that there will be an identical, or sufficiently close, uniformity, as that between the training and evaluation data, in deployment? This uniformity warrants our belief that if a ML model exhibits a decent performance on the held-out evaluation data, it learned a generalisation. Yet the default form of this ML assumption (cf. Vapnik 1999), i.e. independent and identically distributed (i.i.d) phenomena sampled from a fixed distribution, does not furnish an awful lot of leeway. First, without any restrictions imposed on the minimisation of empirical risk the present-day ML models incorporate highly predictive, *yet non-robust*, patterns which are exceedingly likely to exist only in the training-evaluation pair (cf. Ilyas et al. 2019). This makes the attained generalisation (second tier) brittle, the furnished inferences unreliable, and our uniformity assumptions (first tier) unwarranted and severely limited. Second, contrary to our inductive prediction

there might be no uniformity because the held-out evaluation dataset and the next, i.e. deployment environment (target dataset), will be disjoint due to dataset shifts. These shifts are similar to dataset and covariate shifts that might affect any presumed uniformity between training and held-out validation datasets (cf. Quiñonero-Candela et al. 2009). When speaking about dataset shifts further on, we mean changes to the joint distribution of inputs and outputs, covariate shifts on the other hand signify changes in the input distribution only (Quiñonero-Candela et al. 2009, p. XI). Considering finite datasets and disposing of prior uniformity assumptions, the only guarantee that can be provided is relative, i.e. a bound on the difference between the optimal ML model error and a baseline, given the training and evaluation (held-out) dataset (Ben-David 2009). Such a bound will remain loose and could be always superseded by knowledge about any kind of uniformity between the two datasets (cf. Ben-David 2009, p. 82).

Facing dataset shifts and generalisations entailing highly predictive yet non-robust regularities, the only empirical evidence about uniformity can be derived a posteriori from the evaluation (test) set error confidence interval. Such evidence for the uniformity, however, still depends on the human assumption that data generating processes underlying deployment environments/data will be sufficiently like the process which produced the training and evaluation (test) data. For learning to be successful and generalisation (second tier) applicable beyond evaluation (held-out) data, it is necessary to inductively predict or generalise (first tier) that distribution shifts, causing changes to data generating processes in a given domain, are sparse and uniformities abundant. Without this prior human inductive assumption, any generalisation learning becomes a subject to No Free Lunch theorem, which makes its performance (i.e. accuracy) a priori indistinguishable from alternatives (Wolpert 1996).

Under Empirical Risk Minimisation (ERM), which pursues an unconstrained minimisation of error (i.e. incorporating also non-robust regularities) over i.i.d training data (Vapnik 1999), we are left with no assumption-free a priori guarantees that could underpin our inductive prediction regarding uniformity between the evaluation and upcoming data (deployment environments). The generalisation capability of a ML model is measured in terms of its error given as the accuracy gap between the evaluation (held-out test) and training data. Under ERM, any justification of a human inductive inference on the uniformity between the evaluation and deployment environment/data depends on the i.i.d phenomena, as presumably observed in the training data, and 'sparse distribution shifts' assumptions, *not on the generalisation gap of the ML model itself*, provided that training minimised the generalisation gap to the greatest possible extent. Such a dependence can turn into a serious epistemological issue if the uniformity-establishing inductive inferences (predictions, generalisations – first tier) involve *unobservable* elsewheres in space and time where we cannot establish the ground-truth.

The vulnerability stems from an unfavourable signal-to-noise ratio afflicting the reconstructive tasks, which renders the problem ill-posed. Recent efforts to solve such tasks by ML include novel imaging methods designed to harness data acquired by Very Long Baseline Interferometry (VLBI, Barmby 2019). VLBI records radiation emitted by distant celestial bodies by several geographically separated radio telescopes, achieving an angular resolution impossible for a single device (for novel imaging methods e.g. CHIRP [Continuous High-resolution Image Reconstruction using Patch priors], Bouman et al.

2016; Fish et al. 2016). Meeting VLBI and the possibility of processing extended emissions, traditional non-ML reconstructive methods begin to struggle due to sparsity of the data and the level of noise within (ibid.). A possible solution lies in learning a prior from locally available synthetic or natural images, segment the underdetermined image under reconstruction into patches, and retrodict an image in which all the patches are likely under the prior yet aligned with the incomplete and noisy VLBI data (cf. Zoran and Weiss 2011; Bouman et al. 2016; Fish et al. 2016). The resulting image shows a retrodicted spatiotemporal elsewhere, which is furnished by a generalisation acquired from synthetic or local data. The pressing epistemological question remains. What validates our assumption that the uniformity between the training and evaluation (test) data will be preserved also between the latter and target (deployment) environments/data? More precisely, what kind of human inductive prediction/generalisation, underpinning the feasibility of ML inductive risk minimisation, can justify existence of such a uniformity? We will return to this question in Section 3.

Similar epistemological puzzles affect also efforts which use ML models to reach unobservable elsewheres in time. In geosciences, inquiries into past states of the Earth system involve retrodictive inferences furnished by ML generalisations learned from noisy proxies and/or synthetic data (Karpatne et al. 2018). For example, efforts to paint a future picture of the Earth system shaped by anthropogenic changes often involve reconstructions of climates which occurred in the planet's deep past, i.e. paleoclimate reconstructions (e.g. Haywood et al. 2009). As with the unobservable elsewheres in space, elsewheres in time reach the present only as traces deposited in suitable carriers, e.g. deep ice cores, tree rings, corals or lake sediments (Karpatne et al. 2018). The reconstructions are thus empirically underdetermined by paucity of data, ground-truths, and noise affecting the evidence (ibid.). As a result, any ML-based approach to paleoclimate reconstruction will become retrodictive, i.e. inverse and ill-posed. From this setting emerges the epistemically adverse *one-to-many* relation between evidence and many learnable generalisations (second tier), which furnish the inferences to the unobservable past. The relation can be narrowed, and its adverse effects attenuated, only by assuming a uniformity between the present and elsewheres in time. In specific terms, such an assumption depends on uniformitarianism, positing identity between past and present geological processes (Haywood et al. 2009, p. 5), or more deeply on presentism arguing that interpretations of the past can be based on the observable present (cf. Oreskes 2013). In general terms, these uniformity assumptions depend on human inductive predictions (retrodictions) and generalisations about the present.

When using ML models to reach elsewheres in space and time, the domain-specific uniformity assumptions meet presumed uniformities between training and evaluation (held-out) data, and uniformities between the latter and upcoming target/deployment data. Yet the epistemological puzzle persists. What justifies our assumption that the supposed uniformity between the training and evaluation (held-out) data will be preserved also between the latter and target (deployment) environments/data? Section 3 will analyse varieties of human inductive assumptions in ML, above all carefully unpacking their limits, which were suggested already in this subsection. The upshot will entail that despite the existence of domain-specific uniformity assumptions we should not treat ML generalisations about elsewheres as reliable but *merely optimal* regarding the

observable present. Mere optimality stems from the lack of ground-truths that could anchor human inductive predictions/generalisations. In turn, not reliability but mere optimality underlies the uniformity assumptions (first tier) which facilitate ML of generalisations for the retrodictive reconstructions of elsewheres in space or time.

2.2 Non-Benign Environments

The second issue threatening human inductive predictions/generalisations about uniformities that facilitate ML lies in adverse environments. Distribution (dataset) or covariate shifts leading to disjoints instead of uniformities may come not only from reaching elsewheres in space or time but can be also caused by adversaries. While in the former case inductive inferences to uniformity lack ground-truths, in the latter situation we lack guarantees regarding their stability. Two basic types of attacks against existing uniformities can be defined. The first type involves adversaries who choose to attack the uniformity between training and evaluation (held-out) data (cf. Barreno 2010, p. 127). Depending on the practical limits of the adversaries, the outcomes range from altering the uniformity, so that ML generalisations accommodate erroneous inferences, to complete breakdowns of ML caused by dataset or covariate shifts precluding learning and/or evaluation (cf. Laskov and Lippmann 2010). Considering the worst-case scenario in the supervised learning context, training and evaluation (held-out) data would become disjoint, as if sampled from two different distributions (i.e. statistical independence), and labels (the ground-truth) would become independent of the training/evaluation data. Such a scenario breaches the epistemic warrant of any uniformity, as the underlying human inductive predictions/generalisations cannot discern an a priori assumption that would facilitate reliable learning. The consequences become identical to the assumption-free investigation of the off-training set error (Wolpert 1996) because in neither case there are a priori (accuracy) distinctions among the constructible ML models (cf. ibid.).

The first type of attack establishes a basis for security and safety concerns over ML. The second type of attack reveals the instability of ground-truths under certain uniformity assumptions. The instability is caused by adversaries who attack deployed ML models and break the uniformity between evaluation (held-out) and target (deployment) data. The underlying human inductive predictions/generalisations fail because the adversaries exploit ERM's assumption of independent and identically distributed phenomena sampled from a single distribution. Founded on the i.i.d assumption, ERM-based ML cannot distinguish between robust and non-robust regularities. Due to being highly predictive (but unstable), the latter kind of regularity contributes to the minimisation of error of the ML model during training. As a consequence of providing a signal for improving the accuracy of the model, non-robust regularities are not ignored during learning (cf. Ilyas et al. 2019). The crucial point here is that non-robust regularities cannot be covered by human inductive inferences about the uniformity between the evaluation (held-out) and target (deployment) data. The reason for this lies in the features of the data that establish the non-robust regularities. They are incomprehensible to humans and as such their existence will be most likely confined to the training-evaluation data pair (cf. ibid.). Such a conflation between the two types of regularity offers opportunities for creating 'adversarial examples' (i.e. perturbed samples, Goodfellow et al. 2018; Ilyas et al. 2019) designed to collapse the uniformity between the evaluation (held-out) and target

(deployment) data. The adversary seeks to draw the examples so that they either exploit the non-robust regularities, creating samples that manipulate non-robust patterns to malicious ends (e.g. misclassification of the perturbed samples), or draws samples from a distribution that breaks the uniformity altogether, rendering the ML model under attack (entirely) unreliable (cf. ibid.). In both cases, the adversary changes the target (deployment) environment in a way that seeks to negate the human inductive inference about the uniformity which facilitates reliable use of the ML model.

Considering varieties of assumptions in ML, the instability of ground-truths in non-benign environments manifests itself as the agents' attempts to discern a position between the two following extreme scenarios (cf. Gilmer et al. 2018, pp. 17-8). First, the inductive inferences might take the naïve form of predicting adversary-free deployment environments (ibid.). This means fortuitously stable ground-truths, which might purchase, even under ERM's i.i.d assumption, certain predictive mileage regarding uniformity of the evaluation (held-out) and target (deployment) data. However, the instability will quickly return with dataset (distribution) shifts, since under ERM the ML model incorporated non-robust regularities which make the human inductive inferences about the uniformity unreliable. Second, the human inferences might also reflect the worst-case scenario where the target (deployment) environment hosts an unknown number of unbounded adversaries (cf. ibid.). In this situation, it is necessary to expect that the evaluation (test) and target (deployment) data will be disjoint. The possibility of unbounded adversaries renders any prior uniformity assumptions equally uninstructive to guide the selection of a fitting, a priori inductive inference about the ground-truths stability. The situation then becomes subsumed under No Free Lunch theorem.

The upshot of the worst-case scenario entails that despite the possibility of an a priori inference (first tier) to a uniformity between the evaluation and target data we should not treat ML generalisations deployed in adverse environments as reliable. Instead, the human inductive predictions/generalisations to uniformity can be epistemically justified as merely optimal. In situations dominated by inaccessible or unstable ground-truths, optimality turns to be a relational notion. Elsewheres in space and time become furnished with help of ML generalisations which depend on the uniformity between the evaluation and target data. By noise, sparsity, and inaccessibility of the latter, such a uniformity, facilitating ML of generalisations, cannot be epistemically justified as reliable. It is merely optimal in relation to the observed present as captured by the uniformity between the training and evaluation (held-out) data. The human inductive justification of ML models deployed in non-benign environments should be identical. The presumed uniformity between evaluation and target data remains unstable due to bounded and unbounded adversaries. The human inductive prediction/generalisation to uniformity thus lacks future reliability. As a result, the uniformity, and inductive inferences which underlie it, remain merely optimal to our *present* model of the adversaries. The assumptions of the model itself are located somewhere between the two extreme scenarios mentioned above, depending on the prior knowledge the modeller has on the likely capabilities of the adversaries. This knowledge might be of course ill-informed and disproportionately biased to either of the extremes further showing the difference between reliability and optimality justification.

The paper now proceeds to show that the dilemma of justifying inductive inferences, which facilitate ML of generalisations and their uses in target domains, remains an open problem rendered clearer by the concerns over elsewheres and non-benign environments. The dilemma is unpacked by recasting technical assumptions in ML as either human inductive predictions or generalisations about uniformities in data, i.e. about inductive inferences to a low likelihood of distribution (dataset or covariate) shifts.

3 Varieties of Assumptions in Machine Learning

In the present circumstances, machine learning becomes impossible if not interconnected with human inductive inferences. Gauged by positive performances of ML at an ever-growing number of tasks, human intelligence seems capable to discern fitting ML architectures, effective optimisation methods, and other hyper-parameters (cf. Kawaguchi et al. 2019, pp. 15-6). By combining a priori assumptions with the a posteriori evaluation, humans can work out efficacious heuristics which lead to ML models that generalise beyond the sample constituting the training data. A low error on the held-out evaluation data indicates a successful minimisation of empirical (inductive) risk. The human inductive schema, which underpins the heuristics, represents an immediately obvious dimension of the Human-Machine Nexus. Planning for a deployment, especially in situations of inaccessible or unstable ground-truths, the second, arguably more important, dimension of the Nexus comprises inductive inferences to a uniformity between the evaluation (held-out) and target (deployment) data. Without this uniformity, we lose any guarantee that the ML generalisation acquired by learning from the training data still minimises empirical (inductive) risk in target environments. Once deployed, the success of any ML model depends on a particular a priori assumption about the minimisation of empirical risk and its correspondence to the empirical realities of target environments; or more importantly, in the inaccessible/unstable ground-truths situations, the correspondence to our domain-specific suppositions about the targets. At surface level, the correspondence seemingly justifies that an observed uniformity between the training (evidence) and held-out (evaluation) data can serve as a basis for human predictions or generalisations about the relation of the evaluation set to target environments. Any such human inference will lack the epistemic justification if we fail to unpack the assumptions which control the individual flavours of the empirical (inductive) risk minimisation utilised in ML.

The rest of the section pursues this goal by comparing the already mentioned Empirical Risk Minimisation with recently proposed Invariant Risk Minimisation (IRM, Arjovsky et al. 2019). The paper will show that the effort to justify the second anthropic dimension of the Nexus, i.e. human inductive inferences to uniformity, remains unsatisfied. The situation is rendered clearer, yet also worse, for targets based on inaccessible/unstable ground-truths, which comprise elsewheres in space and time and non-benign environments. The issue at hand is a manifestation of Hume's Problem of Induction (1739/1978). Therefore, a recently proposed optimality-based solution (Schurz 2008; 2019) will be used to sketch the epistemic limit of the Human-Machine Nexus.

3.1 Empirical versus Invariant Risk Minimisation

Under ERM, ML models take in patterns existing in the training data, i.e. features, that contribute to minimisation of the loss function which describes the task at hand (cf. Vapnik 1999, pp. 988-9). In the most straightforward, supervised learning case (e.g. pattern recognition), ERM-based models incorporate *any* data features which are correlated with the training data labels (ground-truth) and thus facilitate minimisation of the models' classification loss (cf. ibid.). The aim is to minimise the risk of predicting incorrect class labels for samples coming from the held-out (evaluation) dataset.

Without any restriction as to which data features to include in the model's representations, the attained minimisation of empirical (inductive) risk will depend on two kinds of features. Apart from the robust kind, supposedly aligned with human perception (e.g. cf. Kaur et al. 2019), the model will also incorporate highly predictive yet non-robust features, which are incomprehensible to humans (cf. Ilyas et al. 2019). Not only the latter kind invites adversarial examples (see above; created by flipping the non-robust features to cause erroneous predictions, ibid.), it also makes the human predicted uniformity between the held-out and target data dependent on features which are incomprehensible or, even worse, imperceptible to humans. As a consequence, ERM requires a *fixed* (unknown) distribution generating independent and identically distributed training, evaluation, and target samples, and a *fixed* (unknown) conditional distribution producing labels for those samples (Vapnik 1999, p. 998)[2].

This means that the human inductive inference underpinning ERM presupposes uniformities rather than disjoints. Confronted with dataset and covariate shifts caused by inaccessible/unstable ground truths, ERM may seem divorced from the real-world dynamics. Therefore, ERM can be justified only by understanding the human inductive inference, in this case a prediction, which underpins it. Despite the likely dynamics and variance of target environments, the human predicts that the target data (environment) will comprise solely i.i.d samples coming from the distribution which generated the training and held-out (evaluation) data. Such an inference presupposes a remarkably induction-friendly world where the human becomes justified to count on the extraordinary version of uniformity required by ERM. To reconcile this assumption with real-world dataset/covariate shifts, possibly in target environments lacking accessible/stable ground-truths, the human needs to treat the ML-enabling inductive inference as mere prediction. A prediction to uniformity between the held-out (evaluation) data and a single target environment (data) which comes immediately after the former (held-out) without any change in the underlying data generating process. A stronger inference, e.g. a human inductive generalisation entailing more than one target environment, cannot be justified. Not only because ERM, treated as an inductive principle, does not approximate the observed nature of our world, but also because ERM represents an assumption of local prediction methods (cf. Schurz 2010, pp. 268-9). As such, ERM's reliability depends on an empirical presupposition that cannot change without also voiding ERM's epistemic warrant provided as any justification of the underlying human inductive prediction which facilitates ML (e.g. a uniformity among some material facts, cf. Norton 2003).

[2] ERM depends on further assumptions, which are unrelated to the concern over the uniformity between held-out and target data, and deal with the justification of generalisation bounds for machine learning (Vapnik 1999).

Thus, the Human-Machine Nexus based on ERM becomes ill-suited to be applied in dynamic environments, and especially in those dominated by inaccessible/unstable ground truths. The reason for this lies in the human part of the Nexus. Seeking to satisfy ERM's requirements, the human cannot justify *reliability* of the ML-enabling inference in more than just a few local, and thus predictable, target environments. In a bid to address this issue, Invariant Risk Minimisation was introduced (Arjovsky et al. 2019).

IRM considerably relaxes the ERM's requirement of i.i.d training, evaluation, and target samples drawn from a fixed distribution. IRM presupposes that the training/evaluation samples span a set of diverse target environments, each underlaid by a different data distribution (ibid.). Except for invariant relations between labels and their corresponding phenomena, the training, evaluation, and target environments (data) might be disjoint. The lack of uniformity can be result of disparate conditions in individual environments. The phenomena, whose labels are predicted by an IRM-based ML model, might be captured at different times and/or locations, or could result from interventions shifting the underlying data generating process (e.g. in the case of visual pattern recognition, an object on a variety of non-uniform backgrounds, cf. Arjovsky et al. 2019, p. 2).

Unlike the ERM-based ML, incorporating any data feature correlated with the labels, the IRM-based ML includes only features which remain invariant across the environments and facilitate minimisation of the model's loss function (i.e. classification loss). Compared to ERM, IRM, as a principle of inductive risk minimisation, encourages learning only data features which remain stable across the full range of training environments (Arjovsky et al. 2019, pp. 5-9). The rest needs to be discarded, even if contributing to the risk minimisation locally (i.e. on a particular training data [environment]). Otherwise, the across-environment uniformity would become threatened once again by highly predictive yet non-robust features, evading human comprehension, and thus leaving the human inductive inference to uniformity vulnerable. The IRM-based ML achieves invariance if the trained predictor, a ML model producing labels for the observed phenomena (a classification task), becomes simultaneously optimal for all training environments/data (cf. Arjovsky et al. 2019, p. 5). Practically, this is achieved by finding a favourable trade-off between the predictive optimality across all environments and a local optimality in individual environments separately (IRMv1, cf. ibid.). The predictive optimality is attainable solely in virtue of the *disjoints* (variability) among the individual training environments. Nonetheless, they are still unified by the invariant phenomenon-class label dependency, which holds across the environments and the IRM-based ML models seek to capture it. It is the hope and aim of IRM that by learning invariances, holding across a sufficiently rich landscape of environments, the ML models would be able to robustly generalise about yet unobserved targets containing the captured invariant dependencies.

The human inductive inference underpinning the most expansive interpretation of IRM differs from that of ERM. Relying on diverse sets of training data (environments), and invariances holding across, the human inference to the uniformity between the held-out (evaluation) data and target environments assumes the form of an inductive *generalisation*. Its justification relies on the fact that the supposed uncovered invariance, which underpins the generalisation (first tier), will remain stable in possibly all the so far unobserved target environments containing the phenomenon-class label dependency, i.e. the supposed invariance. Epistemologically, Arjovsky et al. 2019 base their argument for

stability on the supposed causal nature of the phenomenon-class label dependency. While this is far from uncontested (cf. Kilbertus et al. 2018), for justifying the human inductive inference underpinning IRM such a manoeuvre does nothing at all as systematically demonstrated for the first time by Hume (i.e. induction cannot be justified by causality, cf. Schurz 2019, pp. 5-6). Rather, the justification of such a human inductive generalisation depends on the assumption that the observed relative frequency of what is presumably an invariance, i.e. a phenomenon-class label dependency as discerned across the training environments, is close or will uniformly converge to the relative frequency limit, i.e. will not keep randomly oscillating. If by applying IRM to the available training sets, featuring the phenomenon-class label dependency in different empirical settings, the ML model cannot reach the limit, the attained invariance becomes bogus and divorced from the human inductive generalisation. The limit, and thus invariance, cannot be reached unless the observed variance among the training sets supports it. A negative outcome is expected when some of the empirical settings, featuring the phenomenon-class label dependency in radically different ways, are missing from among the training datasets. The incompleteness then lowers the likelihood that the predictor is simultaneously global and local Bayes optimal with respect to the target environments. Whether IRM can in fact compensate for dataset and covariate shifts depends on the human inductive generalisation conjecturing that the phenomenon-class label dependency in the target environments will stay within the bounds of the so far observed variance, hence yielding an invariance. Any attempt to justify such an inductive inference in situations of inaccessible/unstable ground-truths, i.e. elsewheres in space and time and non-benign environments, will face insurmountable obstacles.

The difficulties, however, stem only partially from empirical matters. As in case of ERM, IRM still represents a local prediction strategy. IRM, however, seeks to expand its range by enumerating the phenomenon-class label dependency across various empirical settings with the aim of attaining invariance as the relative frequency limit of the underlying dependency. Therefore, the human generalisation to the uniformity between held-out (evaluation) and target environments can be justified *only* if it is possible to reach, by enumerating the dependency at the object-level, the limit of variance and thus discern an invariance. This approach to justifying inductive inferences follows Reichenbach's pragmatic justification of induction (Reichenbach 1949, pp. 469-82). Here, the core of IRM, that is the enumeration of the phenomenon-class label dependency across diverse environments, is represented by the straight rule (cf. ibid.). If there are discernible relative frequency limits, there are uniformities, the world is thus induction-friendly, and the straight rule leads to justifiable human inductive generalisations. If the opposite is true, human inductive generalisations to uniformities tend fail (cf. ibid.), thus bringing down IRM as well. This train of reasoning underpins the justification of IRM as well, only here it is assumed that relative frequency limits prevail, uniformities are abundant, and IRM-based ML models will thus achieve phenomenon-class label invariances during training. This clearly creates a circle which justifies the IRM-facilitating inductive inference by yet another inductive inference. The straight rule avoided such an outcome by remaining agnostic over the ultimate future reliability of induction (cf. ibid.).

Criticisms of Reichenbach's approach are profound as well as many, and arguably best summarised in Schurz's works on the optimality of meta-induction (Schurz 2008; 2019,

pp. 81-5). Conceding to Hume's objection against a non-circular justification of the future reliability of induction, Schurz showed that at the meta-level, i.e. induction applied to the accessible predictive methods, induction remains optimal (ibid.). Schurz proved that from the present point of view we are justified to prefer induction over other predictive methods because meta-induction always selects the best available prediction method in the long run, and induction performed better than any non-inductive method in the past (ibid.). Schurz thus frustrated the attempt to prove the optimality of induction at the level of *object*-induction (and the reliability of object-induction as well), which Reichenbach implied while responding to some of the objections to the straight rule (ibid.). Because Schurz's justification of the object-level induction can be provided only a posteriori, and only in terms of its present optimality, the a priori future reliability of the human inductive inferences which underpin IRM, and by extension also ERM, cannot be justified. Yet the importance of Schurz's result is unmatched not merely because it provides a non-circular justification of object-induction by proving the optimality of meta-induction in all possible worlds (hence breaking the circularity). Without the optimality of meta-induction, one might argue that ERM/IRM, or in fact any other approach to the ML-based minimisation of predictive risk, could be equally justified by clairvoyant predictions/generalisations. These non- or even anti-inductive inferences would pick up the training environments, *which maximise the uniformity with targets*, by an inscrutable supernatural inspiration.

Besides showing some of the difficulties which affect the epistemic justification of the Human-Machine Nexus, our ability to prove merely the present optimality of induction has also a practical side. The latter comes plainly forward if we consider the Nexus in situations defined by inaccessible/unstable ground-truths.

4 The Contract between the Human and Machine Learning

The present-day Human-Machine Nexus is based on a *fragile contract* (cf. Bottou 2015). After applying some heuristics which enable training of an ML model, its generalisation capability, i.e. the accuracy at a given task, is established by testing the model on an evaluation dataset. If the model reaches an acceptable error, an a posteriori contract between the human and the ML model comes into effect and constitutes an instance of the Nexus. The validity of the contract derives from that specific evaluation set and is guaranteed to hold in the future only if there is a uniformity between the evaluation and target data, i.e. distribution shifts are unlikely occur. By assuming a fixed distribution, this presupposition is problematic in general and perhaps indefensible if the targets comprise inaccessible/unstable ground-truths. These circumstances create a peculiar situation concerning the justification of the contract. We cannot make the contract reliable because human inductive inferences suffering from Hume's Problem can only create ML models which in turn suffer from Wolpert's No Free Lunch theorem. Therefore, the contract is intrinsically fragile, and the Nexus can only produce local prediction methods one at a time. Their justification should be thus based on a contract based on optimality rather than on reliability.

An optimal local predictive method is one which achieves the best possible acquaintance with the presently available states of affairs. Imagine, for a moment, the Human-ML Nexus as a meta-inductive enterprise over globally available objective success records of (empirical/invariant/etc./) risk minimisation (as in Schurz 2008; 2019 only

limited to ML-based local prediction methods). Every a posteriori result of the risk minimisation by an ML model evaluated on a test set, up to now successful or unsuccessful in target environments, would update the success records. To select the presently optimal local predictive methods, humans would simply need to follow Schurz's strategy of attractivity-weighted meta-induction over the success records, which is provably optimal in all possible worlds (ibid.). It could be argued that by following a single experimental paradigm (train/test splits), whose results for ML model/task pairs are published in publicly available accuracy leaderboards (i.e. success records), this is already the case. However, for reasons associated with the availability of records of successful target deployments the whole affair is likely better captured by the local version of meta-induction (Schurz 2012).

From the formal point of view, Schurz's attractivity-weighted meta-induction, producing exponentially weighted averages of the predictions of candidate local methods (Schurz 2019, pp. 138-145), remains universally optimal because it approximates the maximal long-run predictive success of arbitrarily unreliable methods (ibid.). In situations where we cannot gauge the local methods' reliability, following advice of the attractivity-weighted meta-inductivist can never deteriorate but only improve the local method selection process. Even if all candidate local methods were arbitrarily unreliable, which cannot be detected a priori, the meta-induction-based method selection is guaranteed to identify the best among such deceivers, because its attractivity-weighted average predicted in every round approximates the deceivers' maximal success in the long-run and can be used to determine which local method got closest to it. In terms of practical recommendations, establishing ML leaderboards of local methods' success records should be widely encouraged, as the epistemic power of the attractivity-weighted meta-induction increases with a decreasing portion of targets where the success records remain private. Such an empirical practice remains beneficial even if facing inaccessible/unstable ground-truths[3], i.e. noisy, incomplete, proxy data or data generated by adversaries, because the attractivity-weighted meta-induction approximates the maximal long-run predictive success of arbitrarily unreliable methods.

There is, however, a hidden side to optimality-based contracts, which becomes relevant if inaccessible/unstable ground-truths take the centre stage. Optimality-based contracts depend on a posteriori updates of the success records that determine which option, i.e. the combination of a risk minimisation approach, training data, a ML model and its hyperparameters, to follow further considering the changes observed in the target (deployment) environments. Because the updates arrive delayed, optimality-based contracts can work only with the presently available past success records. In case of ML-based reconstructions of elsewheres in space and time, reachable only by sparse, noisy or proxy data, any success depends on local and/or synthetic data, which constitute the uniformity between training and evaluation datasets and thus replace the missing updates from inaccessible ground-truths. In case of non-benign environments, adversaries could induce 'demonic world' conditions characterised by unstable ground-truths where the only uniformity, which can justify a ML model's deployment to the target environment, stems from our present assumptions about the adversaries. In both cases,

[3] A less common example of such a good practice is the leaderboard (scoreboard) of algorithms for VLBI data reconstructions (http://vlbiimaging.csail.mit.edu/).

the success records play a lesser role because the updates to the human assumptions about a training, evaluation, and target dataset uniformity cannot be satisfied by the ground-truths (this of course does not affect the optimality of meta-induction).

The contract between human and machine learning thus transforms from the best possible acquaintance with the available states of affairs, crucially including past success records, to an optimality contract based on the present observable by the humans co-constituting the Nexus. In these circumstances, we lack a transparent or stable access to the ground-truths. The human inductive inferences, which seek to minimise the loss (risk) by maximising the uniformity of ML models with targets, use instead of ground-truths-based updates the second-best thing – our acquaintance with the present, i.e. a reference frame of spatiotemporally situated observers.

Within the bounds of presently available success records, such an incompleteness and indefiniteness of our reference frame (inaccessible and unstable ground-truths respectively) leaves the universal optimality of the attractivity-weighted meta-induction unaffected. The incompleteness might, however, also involve scenarios where our reference frame, and with it also the success records, receive updates that are delayed by physical limits or technical imperfections of the means of information transmission as implied above[4]. In such a case, the optimality-based contract, utilising the attractivity-weighted meta-induction for method selection purposes, might incur increased short-run regrets caused by missing updates to the success records that are yet to arrive. Evaluated with the benefit of *hindsight*, experiencing finite delays of success records updates will translate into a deterioration of the upper bound on short-run losses manifested as a period of decreased uniformity between ML models and targets. During this interim period, human inductive inferences will depend on surrogate data (local, synthetic data or generally data reflecting assumptions about uniformities whose reliability depends on circular justifications) which might fail to maximise the uniformity between ML models and targets, judged from the epistemic position including the delayed updates. Crucially, from the long-run point of view, the contract between human and machine learning based on the attractivity-weighted meta-induction remains universally optimal because it approximates the maximal long-run predictive success of arbitrarily unreliable methods, provided that success records, i.e. leaderboards, are kept public and meticulously updated even on longer time scales.

5 Conclusion

The contract which constitutes the Human-Machine Nexus is fragile because the inductive inferences, which underpin varieties of assumptions in ML, can be justified as merely optimal. The optimality-based contract refers to constant inflow of updates, which, by chronicling failures and successes of the previous attempts, permit maximising the uniformity between ML models and target environments to the best of our present knowledge, crucially including past success records. In case of inaccessible/unstable ground-truths, the contract shifts and begins to depend on our acquaintance with the present, or more generally on reference frames of spatiotemporally situated observers.

[4] We are indebted to a referee for the journal who suggested this addition and pointed out yet another valuable perspective from which to appreciate the universal long-run optimality of Schurz's attractivity-weighted meta-induction.

This version of optimality-based contract invites a kind of presentism. Under its aegis, reconstructions of distant elsewheres and resilient ML models can be obtained if we replace the inaccessible/unstable ground-truths with present information stemming from our local environment. Yet the Nexus should help us to extend ourselves and to do so in a resilient manner. By augmenting our cognition with ML, we seek to reach distant epistemic locations and become resilient in the face of adversity. It is perhaps a fitting paradox that when facing inaccessible/unstable ground-truths we transform ourselves into 'resilient super-observers' constituted by the Human-Machine Nexus. This is fine because ML has become our best chance to move forward and satisfy our epistemic ambitions, common and scientific alike. While pursuing the Nexus further we should, however, never lose sight of the fragile contract which justifies the varieties of human assumptions enabling ML. For now, moving forward is likely to remain a matter of careful interplays in the Nexus of human and machine learning. And for that, the epistemology of machine learning will become a useful instrument.

Acknowledgments

We are grateful to our reviewers for the helpful feedback as well as to the editors at ACM Computing Surveys.
This study was supported by the Charles University Research Programme "Progres" Q18 – Social Sciences: From Multidisciplinarity to Interdisciplinarity.

References

<BIBL>
Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565v2 [cs.AI]. Retrieved from https://arxiv.org/abs/1606.06565v2
Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. arXiv:1907.02893v2 [stat.ML]. Retrieved from https://arxiv.org/abs/1907.02893v2
Pauline Barmby. 2019. Astronomical observations: a guide for allied researchers. *The Open Journal of Astrophysics*. DOI: https://doi.org/10.21105/astro.1812.07963
Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and machine learning: Limitations and Opportunities*. Manuscript in preparation. Retrieved from https://fairmlbook.org/
Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. 2010. The security of machine learning. *Machine Learning* 81 (Nov. 2010), 121-48. DOI: https://doi.org/10.1007/s10994-010-5188-5
Shai Ben-David. 2009. On the Training/Test Distributions Gap: A Data Representation Learning Framework. In *Dataset Shift in Machine Learning*, Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (Eds.). MIT Press, Cambridge, MA, 73-84.
Nick Bostrom. 2014. *Superintelligence, Path, Dangers, Strategies.* Oxford University Press, Oxford.
Léon Bottou. 2015. Two Big Challenges in Machine Learning. A Keynote at 32nd International Conference on Machine Learning (ICML '2015). Retrieved from https://leon.bottou.org/talks/2challenges

Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, and William T. Freeman. 2016. Computational Imaging for VLBI Image Reconstruction. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, June 27-30 2016, Las Vegas, NV. Curran Associates, Inc., 913-922. DOI: https://doi.org/10.1109/CVPR.2016.105

Nicolo Cesa-Bianchi and Gabor Lugosi. 2006. Prediction, Learning, and Games. Cambridge University Press, Cambridge.

David Corfield, Bernhard Schölkopf, and Vladimir Vapnik. 2009. Falsificationism and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions. *Journal for General Philosophy of Science* 40 (Jul. 2009), 51-58. DOI: https://doi.org/10.1007/s10838-009-9091-3

David Corfield. 2010. Varieties of Justification in Machine Learning. *Minds & Machines* 20 (Jul. 2010), 291-301. DOI: https://doi.org/10.1007/s11023-010-9191-1

Virginia Dignum. 2018. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20, 1 (Mar. 2018), 1-3. DOI: https://doi.org/10.1007/s10676-018-9450-z

Vincent L. Fish, Kazunori Akiyama, Katherine L. Bouman, Andrew A. Chael, Michael D. Johnson, Sheperd S. Doeleman, Lindy Blackburn, John F. C. Wardle, and William T. Freeman. 2016. Observing—and Imaging—Active Galactic Nuclei with the Event Horizon Telescope. *Galaxies* 4, 4 (2016), 7-11. DOI: https://doi.org/10.3390/galaxies4040054

Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. 2018. Motivating the Rules of the Game for Adversarial Example Research. arXiv:1807.06732v2 [cs.LG]. Retrieved from https://arxiv.org/abs/1807.06732v2

Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making Machine Learning Robust Against Adversarial Inputs. *Communications of the ACM* 61, 7 (Jul. 2018), 56-66. DOI: https://doi.org/10.1145/3134599

Gilbert Harman and Sanjeev Kulkarni. 2007. *Reliable reasoning: induction and statistical learning theory*. MIT Press, Cambridge, MA.

Alan M. Haywood, Harry J. Dowsett, Paul J. Valdes, Daniel J. Lunt, Jane E. Francis, and Bruce W. Sellwood. 2009. Introduction: Pliocene Climate, Processes and Problems. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 367, 1886 (2009), 3-17. DOI: https://doi.org/10.1098/rsta.2008.0205

David Hume. 1739/1978. *A Treatise on Human Nature. Book I: On Human Understanding*. Oxford University Press, Oxford.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS '2019)*, December 8-14, 2019, Vancouver, Canada. Curran Associates, Inc., 125-136.

Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan A. Babaie, and Vipin Kumar. 2018. Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering* 31, 8 (Aug. 2018), 1544-54. DOI: https://doi.org/10.1109/TKDE.2018.2861006

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. 2019. Are Perceptually-Aligned Gradients a General Property of Robust Classifiers? In *the Science meets Engineering of Deep Learning NeurIPS 2019 Workshop*. Retrieved from https://arxiv.org/abs/1910.08640v2

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2019. Generalization in Deep Learning. arXiv:1710.05468v5 [stat.ML]. Retrieved from https://arxiv.org/abs/1710.05468v5

Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. 2018. Generalization in anti-causal learning. In *the Critiquing and Correcting Trends in Machine Learning NeurIPS 2018 Workshop*. Retrieved from https://arxiv.org/abs/1812.00524

Pavel Laskov and Richard Lippmann. 2010. Machine learning in adversarial environments. *Machine Learning* 81 (Nov. 2010), 115-19. DOI: https://doi.org/10.1007/s10994-010-5207-6

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521 (May 2015), 436-44. DOI: https://doi.org/10.1038/nature14539

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

Vaishnavh Nagarajan and J. Zico Kolter. 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS '2019)*, December 8-14, 2019, Vancouver, Canada. Curran Associates, Inc., 11615-11626.

John D. Norton. 2003. A Material Theory of Induction. *Philosophy of Science* 70, 4 (Oct. 2003), 647-70. DOI: https://doi.org/10.1086/378858

Naomi Oreskes. 2013. Why I am a Presentist. *Science in Context* 26, 4 (Dec. 2013), 595-609. DOI: https://doi.org/10.1017/S026988971300029X

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (Eds.). 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA.

Hans Reichenbach. 1949. *The Theory of Probability*. University of California Press, Berkeley, CA.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (Jan. 2015), 85-117. DOI: https://doi.org/10.1016/j.neunet.2014.09.003

Gerhard Schurz. 2008. The Meta-inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem. *Philosophy of Science* 75, 3 (Jul. 2008), 278-305. DOI: https://doi.org/10.1086/592550

Gerhard Schurz. 2010. Local, General and Universal Prediction Methods: A Game Theoretical Approach to the Problem of Induction. In *EPSA Epistemology and Methodology of Science*, Mauricio Suárez, Mauro Dorato, and Miklós Rédei (Eds.). Springer, Dordrecht, 267-78. DOI: https://doi.org/10.1007/978-90-481-3263-8_23

Gerhard Schurz. 2012. Meta-Induction in Epistemic Networks and the Social Spread of Knowledge. *Episteme* 9, 2 (Jun. 2012), 151-70. DOI: https://doi.org/10.1017/epi.2012.6

Gerhard Schurz. 2017. Optimality justifications: new foundations for foundation-oriented epistemology. *Synthese* 195 (Sep. 2018), 3877-97. DOI: https://doi.org/10.1007/s11229-017-1363-6

Gerhard Schurz. 2019. *Hume's Problem Solved: The Optimality of Meta-Induction*. MIT Press, Cambridge, MA.

Gerhard Schurz and Paul Thorn. 2020. The material theory of object-induction and the universal optimality of meta-induction: Two complementary accounts. *Studies in History and Philosophy of Science Part A* 82 (Aug. 2020), 88-93. DOI: https://doi.org/10.1016/j.shpsa.2019.11.001

Kaj Sotala and Roman V. Yampolskiy. 2015. Responses to catastrophic AGI risk: a survey. *Physica Scripta* 90, 1 (Jan. 2015), 1-33. DOI: https://doi.org/10.1088/0031-8949/90/1/018001

Petr Spelda. 2018. Machine learning, inductive reasoning, and reliability of generalisations. *AI & Society* 35 (Mar. 2020), 29-37. DOI: https://doi.org/10.1007/s00146-018-0860-6

Vladimir Vapnik. 1999. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks* 10, 5 (Sep. 1999), 988-99. DOI: https://doi.org/10.1109/72.788640

Ulrike von Luxburg and Bernhard Schölkopf. 2011. Statistical Learning Theory: Models, Concepts, and Results. In Handbook of the History of Logic, Vol. 10: *Inductive Logic*, Dov M. Gabbay, Stephan Hartmann, and John Woods (Eds.). North Holland, Kidlington, 651-706. DOI: https://doi.org/10.1016/B978-0-444-52936-7.50016-1

Vladimir Vovk. 2015. Comment: The Two Styles of VC Bounds. In *Measures of Complexity Festschrift for Alexey Chervonenkis*, Vladimir Vovk, Harris Papadopoulos, Alexander Gammerman (Eds.). Springer, Cham, 161-64. DOI: https://doi.org/10.1007/978-3-319-21852-6_11

David Wolpert. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* 8, 7 (Oct. 1996), 1341-90. DOI: https://doi.org/10.1162/neco.1996.8.7.1341

Roman V. Yampolskiy (Ed.). 2019. *Artificial Intelligence Safety and Security*. CRC Press, Boca Raton, FL.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *the 5th International Conference on Learning Representations (ICLR '2017)*, April 24-26, 2017, Toulon, France.

Daniel Zoran and Yair Weiss. 2011. From learning models of natural image patches to whole image restoration. In *Proceedings of the 2011 International Conference on Computer Vision (ICCV'11)*, November 6-13, 2011, Barcelona, Spain. Curran Associates, Inc., 479-486. DOI: https://doi.org/10.1109/ICCV.2011.6126278

</BIBL>