The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning

Jessica Hullman[1,4], Sayash Kapoor[2], Priyanka Nanayakkara[1], Andrew Gelman[3], Arvind Narayanan[2*]

ABSTRACT

Recent concerns that machine learning (ML) may be facing a reproducibility and replication crisis suggest that some published claims in ML research cannot be taken at face value. These concerns inspire analogies to the replication crisis affecting the social and medical sciences, as well as calls for greater integration of statistical approaches to causal inference and predictive modeling. A deeper understanding of what reproducibility concerns in researchin supervised ML have in common with the replication crisis in experimental science can put the new concerns in perspective, and help researchers avoid "the worst of both worlds" that can emerge when ML researchers begin borrowing methodologies from explanatory modeling without understanding their limitations, and vice versa. We contribute a comparative analysis of concerns about inductive learning that arise in different stages of the modeling pipeline in causal attribution as exemplified in psychology versus predictive modeling as exemplified by ML. We identify themes that re-occur in reform discussions like overreliance on asymptotic theory and non-credible beliefs about real-world data generating processes. We argue that in both fields, claims from learning are implied to generalize outside the specific environment studied (e.g., the input dataset or subject sample, modeling implementation, etc.) but are often impossible to refute due to forms of underspecification. In particular, many errors being acknowledged in ML expose cracks in long-held beliefs that optimizing predictive accuracy using huge datasets absolves one from having to make assumptions about the underlying data generating process. We conclude by discussing rhetorical risks like error misdiagnosis that arise in times of methodological uncertainty.

CCS CONCEPTS

• **Computing methodologies → Learning paradigms**; **Supervised learning**.

KEYWORDS

Machine learning, replication, science reform, generalizability.

1 INTRODUCTION

The replication crisis in psychology and the social and medical sciences has spread to a general concern about scientific claims that are based on statistical significance. Similar attention has recently been drawn to replication challenges regarding empirical claims in artificial intelligence (AI) and machine learning (ML). There are direct concerns about *reproducibility*—published results cannot be reproduced using the same software and data due to unavailable tuning parameters, random seeds, and other configuration settings or computational infrastructure that are not available to outsiders—*replication*—where re-implementing described methods does not produce the same results due to unacknowledged dependencies, such as specific implementations, and—*robustness*, where methods may work well under certain conditions but fail when applied to new problems or in the world, where vulnerability to adversarial manipulations may be costly. For example, the identification of examples by which computer vision models could be tricked into misclassification by manipulations not visible to the human eye [190] has inspired subsequent research proposing a variety of explanations for the apparent brittleness of performance (e.g., [52, 74, 90]). Terms like "alchemy" [115] and "graduate student descent" are used to describe how researchers combine optimizations to often opaque parameters to achieve performance benchmarks. Model performance evaluations are conducted without acknowledging sources of error [2, 93, 130] and can involve data filtering decisions that impact achievable accuracy [129, 130].

Some amount of replication failure is inevitable: the nature of empirical research is to try out ideas that may work in some settings but not others. When claims are published, uncertainty about generalizability is inherent. However, once systemic problems are recognized, published claims should be discounted in some way—especially when they cannot be externally reproduced [40, 80, 117].

In the social sciences, while fundamental issues with conventional approaches to inference were discussed by psychologists at least as early as the 1960s [48, 139], the last decade saw a shift from perceptions that reformers were engaging in standard forms of scientific methods criticism to perceptions that entire fields are built on unwarranted conclusions. Critics demonstrated how motivated researchers can obtain false positives under various conditions [70, 71, 180] and that many published conclusions about human behavior in psychology research cannot be replicated [65, 152]. These revelations spur deep questions about what constitutes necessary conditions for science, how to resolve uncertainty about published claims, and how to shift incentives in a field.

Similar revelations and questions could benefit AI and ML research. However, these communities are currently at risk of failing to grasp important lessons in emerging critiques. If published at all, arguments criticizing the reproducibility or replicability of conventional practice may not appear in mainstream venues and may seem difficult to relate to one another. Some may see the concerns as minor in comparison to the accuracy gains of deep learning models.

Learning failures in both causal attribution (often associated with explanation goals as in psychology research) and predictive modeling (as in AI and ML research) have also inspired proposals for integrative modeling approaches that combine aspects of both [107, 108, 210]. On the surface, a psychology researcher trying to understand how a person's traits or thought patterns impact their behavior has little in common with an ML researcher trying to

[*] First author is corresponding author.

1. Northwestern University. {jhullman@, priyankan@u.}northwestern.edu
2. Princeton University. {sayashk@, arvindn@cs.}princeton.edu
3. Columbia University. gelman@stat.columbia.edu
4. Microsoft Research New York City.

develop the most accurate possible model to label images or predict customer churn. However, both fields study dynamic systems with complex, emergent behavior [18], where explanation and prediction both have important roles. Despite well-known differences in their goals (e.g., [36]), psychology researchers doing explanatory modeling often assume their models have predictive validity [210], and AI and ML researchers implicitly compare model behaviors to their beliefs about the world [74, 116]. As researchers trained in one type of modeling adopt conventions of the other, it is important they grasp associated limitations of the methods. Otherwise integrated approaches might lead to “the worst of both worlds.”

As a step toward synthesizing concerns with research claims in AI and ML with more established discussions of reproducibility, we contribute a comparative analysis of concerns about inductive learning that arise in different stages of the modeling pipeline in causal attribution-focused versus predictive modeling. To ground our discussion, we use examples from experimental social psychology, which like ML relies on data reflecting human behavior and uses controlled comparisons to produce claims. In ML, we focus on empirical research in supervised discriminative learning (i.e., classification) methods, including deep neural nets (DNNs) that encapsulate many concerns in the burgeoning reform literature. Our analysis synthesizes formal and informal arguments made in hundreds of papers we collected through online search, citation tracing, and our involvement in events and scholarly discussions on replication and reproducibility over multiple years.

Despite commonly-touted differences between the two fields, we observe many points where concerns stem from similar types of oversights. These include overreliance on irrelevant theory, underspecification of learning goals, non-credible beliefs about real-world data generating processes, overconfidence based in conventional faith in certain procedures (e.g., randomization, test/train splits), and tendencies to engage in dichotomous reasoning about empirical results. We argue that in both fields, claims from learning are implied to generalize outside the specific environment studied (e.g., the input dataset or subject sample, modeling implementation, etc.) but are often impossible to refute due to forms of underspecification. We conclude that many of the errors recently discussed in ML expose the cracks in long held beliefs that optimizing predictive accuracy using huge datasets absolves from having to consider a true data generating process or formally represent uncertainty in performance claims. We conclude with high-level recommendations for researchers in both fields.

2 BACKGROUND

2.1 Anatomy of a learning process

An idealized learning process begins with the formulation of **goals** (including scientific goals such as understanding what factors influence a particular human behavior, engineering goals such as constructing a better model for machine translation, or even policy goals such as estimating effectiveness among different categories of patients) and **hypotheses**. These are not necessarily statistical “hypotheses”; rather, a hypothesis could be that a certain thinking pattern increases the chances of a behavior, or that a certain technical innovation will lead to a better translation system, or that a treatment will be more effective among men than women. Goals and hypotheses lead to steps of **data collection and preparation**. Researchers specify an observational process to collect information about the latent phenomena of interest from the environment. An observational probe is used to induce explicit observations thought to be sensitive to the target phenomena. For example, psychologists design human subjects experiments around interventions thought to interact with the target phenomena. AI and ML researchers often make use of datasets generated from human produced media and signals of behavior, in the form of digital traces.

An observational process becomes a model by making assumptions about what the observed data represent, namely realizations of random variables with regular variation. The observational model is defined by a choice of **model representation**, i.e., the model class or functional form that consequently specifies a space of data generating processes (DGPs, i.e., fitted functions) that might have produced the data. This might be a multiple linear regression functional form in social psychology, or a DNN architecture in ML defined by the configuration of network parameters (e.g., arrangement into convolutions, activation functions, pooling method, etc.). Because quantifying and searching *all* data generating processes implied by probability distributions over the observation space tends to be prohibitively complex, learning pipelines typically consider a subset or “small world” of model configurations [25], called the hypothesis space of the learner in ML. **Model selection** or model-based inference describes how the space of model configurations is traversed to find a best fit model that is most consistent with the data. This involves defining an objective or loss function measuring the difference between the ground truth observed outcome for an input and the predicted outcome of a parameterized model configuration (e.g., squared error), as well as an optimization method for searching the space of model configurations to find the fitted function that minimizes loss (e.g., gradient descent, adaptive optimization algorithms, analytical solutions like MLE, etc.).

An **evaluation** may follow to validate the usefulness of what is learned, such as relative to alternative model fits or learning pipelines. Evaluation metrics such as explained variance or log loss can be used to summarize overall usefulness of a fitted function. Evaluation metrics may sometimes be implicit, such as when the usefulness of a fitted model is evaluated relative to one’s hypotheses about the data generating process. The learning process culminates in **communication** of claims in research papers.

2.2 Goals of learning in social psychology versus machine learning

Social psychology. A primary goal in empirical psychology is to describe the causal underpinnings of human behavior [139, 179, 210]. Researchers identify hypotheses representing predictions about variables that constitute observed data. Often these constitute “weak theories” [140], predicting a directional difference or association between variables but not the size of the effect. They design observational processes to gather data for testing hypotheses, typically controlled human-subjects experiments that record the thoughts, emotions, or behavior of subjects, under different conditions thought to interact with the latent phenomena of interest. The approximating functions that researchers learn from these observations (often low dimensional linear regressions) are thought to capture key structure in the latent psychological phenomena. Claims about cause and effect hinge on interpreting the parameter values of the fitted function in light of hypotheses and their sampling variation within a statistical testing framework. A function is commonly deemed worthy of interest if it is below a false-positive rate defined in the Neyman-Pearson framework, typically $\alpha = 0.05$. Direct claims take the form of statements about novel statistically significant causal attributions, and have been called “stylized facts” [79, 106] implied by authors to be generally true about human behavior. For example, thinking about old age induces old-like behavior [12].

Machine learning. A primary goal in supervised ML research is to facilitate the learning of functions which achieve high predictive accuracy in tasks like classification. Researchers hypothesize procedures or abstractions that may improve the state-of-the-art in subareas (e.g., NLP, vision), which is captured by benchmarks: abstractly defined tasks (e.g., image classification, machine translation) instantiated with learning problems consisting of datasets (input, output pairs) and an associated evaluation metric to be used as a scoring function (e.g., accuracy) [130]. Standard methods like using a train-test split and cross validation are designed to ensure good predictive performance of a fitted model on unseen data.

Claims made in empirical research papers typically report performance of a new learner (i.e., fitted model) on benchmarks, compared to baselines representing the prior state-of-the-art. Formal proofs of the statistical properties of new methods are also common.

3 THREATS TO LEARNING IN SOCIAL PSYCHOLOGY AND MACHINE LEARNING

We describe threats to valid learning according to whether they involve data selection and preparation, model development (including defining a representation and model selection and evaluation approach), and communication of results in a research paper.

3.1 Data collection and preparation

Social psychology. *Problems of high measurement error and variation relative to signal, unacknowledged flexibility in defining data inputs, and underspecified or non-representative subject samples are frequently associated with validity issues in psychology. Overlooking the importance of stimuli sampling and other "design freedoms" can similarly contribute error.*

Claims are threatened by the use of small samples combined with noisy measurements, the effects of which are poorly understood by researchers. For example, one pervasive belief is that if an experiment registers a “statistically significant” effect on a small sample, then that effect will necessarily remain significant with a larger sample [40, 180]. In reality, with a lower powered study, not only is there a lower probability of finding a true effect of a given size, there is a lower probability that an observed effect which passes a significance threshold actually reflects a true effect that will appear under replication [40]. Under low power, estimates of observed effects will tend to reflect sampling error, deriving from the limited size of the sample relative to a target population, and forms of measurement error [133], such as random variation due to noise in taking measurements that produces a difference between observed and true values. Studies are “dead on arrival” when standard error due to measurement and sampling variation is large relative any plausible effect size [86].

Inherent flexibility in how a researcher specifies an analysis presents another threat to inference. A “researcher degrees of freedom” or “garden of forking paths” metaphor [83, 180] suggests that given human tendencies toward self-serving interpretations of ambiguous evidence (e.g., [10, 55] as cited in [180]), researchers are likely to draw conclusions that verify their hypotheses. Given an outcome variable, set of covariates and functional form, an analyst may bias results toward a preferred conclusion during data preparation by selecting data transformations and outlier removal processes after observing the data, without necessarily recognizing they are doing anything wrong. A researcher can also often choose between different outcome and predictor variables and ways of operationalizing them, leading to more unacknowledged flexibility. More broadly, when a researcher can tweak the design of experiment conditions with feedback through pilot experiments via the design of stimuli, instructions, and elicitation instruments, they may gravitate toward designs that exaggerate effects in some conditions. Overlooking the importance of stimuli

Scholars have pointed to irreproducibility of psychology study results based on learning occurring on non-representative samples of a target population, such as convenience samples of university students from Western educated industrialized rich democratic (WEIRD) countries [104]. As researchers have become more accustomed to the importance of statistical power and representative samples, online recruitment of participants in social psychology [172] increases. However, it is unclear that sample homogeneity is addressed by online samples [44] and this trend has led to greater use of self-reported measures [172] that contribute additional noise. More generally, failure to recognize the implications of non-random sampling can lead to a “big data paradox” of overconfidence as sample size increases [142]. Similarly fundamental but often overlooked issues concern how psychologists often leave the target population unspecified [88] and fail to consider the importance of sampling stimuli as well as subjects [87, 204, 209], calling into question of what is being learned at all.

Machine learning. *In ML research, standardization of benchmarks and the prohibitive cost of amassing large datasets means that data selection often means selecting among existing datasets [101, 188], typically obtained through crowdsourced annotation and web-scale data (e.g., [58, 127]). We see several points of overlap with SP data issues related to flexibility in data transformation, use of non-representative samples, and underspecification of the population captured in data, but also a heavier emphasis on forms of non-random measurement error and a unique, normative viewpoint on how model predictions can perpetuate real world stereotypes.*

First, scholars have begun to point to analogous concerns to psychology in recent acknowledgement of flexibility in data transformation, such as in filtering data in ways that simplify a prediction problem (e.g., removing translation artifacts in machine translation to improve prediction accuracy [129] as cited in [130]).

We also see emphasis on non-representative samples. One commonly cited root cause is violation of the assumption that the training (or “development”) distribution from which the data used for learning are presumed to be randomly drawn is the same as the use (or “deployment”) distribution from which random samples will be drawn when applying the model in practice [13, 156, 189]. Development data that underrepresent some parts of the input space of an ML algorithm, leading to higher error rates for less represented instances in the input space (e.g., [39, 155, 213]) has been termed "representation bias" [189]. Suresh and Guttag [189] define this bias as a positive value for a measure of divergence between the probability distribution over the input space and the true distribution, noting that it can occur simply as a result of random sampling from a distribution where some groups are in the minority. Others point to the potential for overlooked errors in the labeling process, which is often left undescribed in research papers [73], to lead to overfitting even in the absence of other types of noise [35, 151], and the way that data preparation can be lossy whenever majority-rule is used to construct ground truth without preserving information about label distributions [54, 92].

However, we see a greater emphasis on measurement error in the form of systematic bias in collected measurements that threatens construct validity, whether the measurement is actually capturing the intended concept. "Measurement bias" [189] has been used to refer to differential measurement error [198], where a measurement proxy is generated differently across groups due to differing granularity or quality of data across groups, or reduction of complex target category (e.g., academic success) to a small number of proxies that favor certain groups over others (e.g., [126] as cited in [189]). Jacobs and Wallach [118] attribute many misleading claims in the fairness literature in ML to unacknowledged mismatches between unobservable theoretical constructs in ML applications (e.g., risk of recidivism, patient benefit) and the measurement proxies that researchers often tend to assume capture them, and suggest the use of latent variable models to formally specify assumptions.

Perhaps the most novel concerns related to measurement bias in ML relative to SP occur when biased input data are used to train a model and contribute to undesirable social norms. Data may record historical biases [189] (e.g., training a model to recognize successful applicants on data where women were admitted less due to bias). "Harms of representation" [1, 50] refers to how model predictions can reinforce potentially harmful stereotypes when trained on data exhibiting bias. For example, returning pictures of only white males on a Google search for CEO reinforces notions that other groups are not as appropriate for CEO positions [124]. The fact that ML is often intended for prescriptive use in the world, rather than descriptive use as in psychology research helps explain these concerns as well as the emphasis on systematic measurement error.

Finally, data concerns in ML increasingly refer to forms of underspecification of population characteristics and underacknowledgment of the constructed nature of data in the convention of taking data as given [19, 73, 114, 174]. These concerns also imply that real world harms may result from practices that extract away the subjective judgments, biases, and contingent contexts involved in dataset production [156].

3.2 Model representation

Learning from data requires selecting, at least implicitly, a formal representation that defines what functions can be learned.

Social psychology. *Beliefs in the potential for unbiased studies lead psychology researchers to overconfidently interpret model estimates. Researchers commonly overlook the importance that the small world of model configurations they explore captures or well approximates the true DGP for valid inference, hold unrealistic views about the separability of large effects in the world, and tend to ignore the value of incorporating prior knowledge in modeling.*

Consider how psychology researchers most commonly model their data using simple measures of correlation and linear parametric models [29, 30] as cited in [199], then check the significance of certain model coefficients in order to decide what has been "discovered." In designing an observational process to gather information about a latent psychological phenomena of interest, then using the observations to fit a model (Section 2.1), researchers implicitly assume that there is a true data generating process that exactly captures how the target phenomena arises as a function of other factors thought to influence it. Once an observational model is defined, inference is confined to the mathematical narratives represented by these functions [84]. However, the validity of claims made about causal effects by following this process depend upon judicious choices about how to represent structure in the true DGP in the constrained small world model space, which psychology researchers often overlook [199].

A first complication arises from the fact that inference is more straightforward when the true DGP is included in the small world of configurations under consideration [24]. However, the sorts of human behaviors psychologists tend to target are thought to be conceptualizable but too complicated to specify explicitly, or not even conceptualizable [24, 199]. Under these conditions, the validity of conventional interpretations of fitted models depends on the observational model faithfully approximating the true DGP [84].

However, this is not the case when a model is structurally misspecified, meaning the fitted models do not adequately capture the true causal structure and/or the functional form of the relationships between variables in the true DGP [199]. For example, if the DGP in a psychology study can be described as a weighted sum of the set of input variables that are represented in the chosen functional form, and all of these predictors are exogenous (i.e., completely independent), then parameters estimated using ordinary least squares can be interpreted according to convention as information about the target phenomena (e.g., comparing two items that differ by one unit in predictor x while being the same in all other predictors will differ in y by θ, on average). However, when the true DGP is more complex than the functional form, as is usually the case, the choice of which potential confounding variables one measures and includes in the regression equation becomes important. Not including variables that influence a regressor and the outcome [6] or including variables that could in principle be affected by experimental manipulations (and hence represent outcome variables themselves [49]) cause the conventional interpretation of the fitted parameter values not to hold. However, researchers seldom acknowledge these limitations.

Researchers also often decide what experimental design to use based on a preference for simpler models. Perhaps the most common example is a preference for between-subject designs based on beliefs that they provide cleaner estimates based on their asymptotic properties: as the size of the (random) sample increases toward the population size, a between-subjects design provides a simpler procedure for estimating average treatment effects relative to a within-subjects design, which requires estimating various carry-over effects between treatments experienced by the same individual

	Social Psychology	**Machine Learning**
Data selection and preparation	○ Measurement error [14, 59, 133]	○ Differential measurement error [39, 155, 189, 213]; unmodeled measurement error [118, 126]
		○ Label errors [35, 151] and disagreement [54, 92]
	○ Data transformations decided after seeing results [83, 180]	○ Data transformations decided after checking performance [129]
	○ Non-representative [104, 142] or underdefined subject samples [88]; insufficient stimuli sampling [87, 204, 209]	○ Underrepresentation of portions of input space in training data [13, 156, 189]
	○ Small samples and noisy measurements (low power) leading to biased estimates [40]	○ Input data too huge to understand [19, 156]
Model representation	○ Overreliance on models and designs with good asymptotic guarantees [150]	○ Overreliance on asymptotic (worst-case) guarantees [62]
	○ No explicit representation of prior/domain knowledge [75, 84]	○ Underspecification of desired inductive biases [52, 116]; failure to prevent shortcut learning [74]
	○ Inappropriate fixed effect expectations [47, 77, 209] in light of crud factor [140, 153]; belief in many nudging factors with large consistent effects on outcome [196]	○ Inappropriate i.i.d. assumption in light of real-world nonstationarity [28, 186, 206]
		○ Reliance on fine-tuning/foundation models for which hyperparameter tuning is opaque [61, 207]
	○ Unacknowledged multiplicity of solutions [210]	○ Convergence in architectures around large models [19, 31, 187]
	○ Unacknowledged structural misspecification [136, 199]	
Model selection and evaluation	○ Implicit optimization for statistical significance [80, 82, 83, 91, 113, 128]	○ Implicit optimization to beat SOTA [107, 176]
	○ Inference as black box [88, 128, 203]; Not motivating choice of estimator or optimization for particular inference goal [22, 201]	○ Overlooked sensitivity of optimizer performance to hyperparameters [34, 45, 175]; computational budget [194]
		○ Presence of implementation variation [130] and tricks [5, 103]
	○ Misunderstanding/misusing ideas of statistical significance [76, 96, 112, 113, 200]	○ Misuse of cross validation [23, 43, 102, 102]
		○ Optimism of cross validation [68, 137]
	○ Multiple comparisons problem [81]	○ Loss metric misalignment [111]
		○ Not comparing to simpler baselines [176] or priors [95]
Communication of claims	○ Unwarranted speculation about what evidence a p value provides [192]	○ Unwarranted speculation about causes behind results [123, 130, 132]
	○ Overgenerationalization (i.e., beyond studied population) [57, 99, 167, 182, 209]	○ Implying equivalence of learning problems and human performance on a task [123, 130, 132]
	○ Unavailable data and code [78, 109, 144]	○ Lack of dataset documentation [19, 73, 156]
		○ Inaccessible data, code, computational resources [93, 171, 185]
	○ Not acknowledging having explored multiple analyses conditioned on data [81, 180]	○ Not reporting implementation conditions/sources of variance [132, 176]
	○ Inaccurate descriptions of what p values mean [3, 27, 85, 192]	○ Underpowered performance comparisons [2, 34]; ignoring sampling error [2, 130, 162]

Table 1: Overview of learning concerns, ordered to emphasize similarities across social psychology and ML.

[150]. However, researchers tend to overlook limitations when sample size is smaller, like how between-subjects designs do not control for variation between people, which can be quite large, and can lead to biased estimates of average treatment effects if the treatment interacts with "background variables" associated with individual differences, different contexts, or different situations [136, 150].

More generally, researchers tend to expect constant effects rather than assuming effects will vary across people or contexts [77]. This can manifest, for example, as model specifications that ignore the importance of modeling variation in stimuli and other experimental conditions as well as subjects [209] (e.g., a "fixed effect fallacy" [47]).

Tendencies to overlook important sources of variation in modeling are implied by Meehl's conception of the "crud factor" [140, 153], which emphasizes how causal attribution based on constrained model spaces used to approximate a highly complex true DGP is fundamentally challenged by the prevalence of "real and replicable correlations" reflecting "true, but complex, multivariate and non-theorized causal relationships" between all variables [153]. Problems arise when questionable beliefs about reality lead researchers to overlook forms of model misspecification. For example, a tendency toward reporting model fits suggesting that novel yet seemingly trivial "nudging" factors (e.g., whether or not someone is menstruating or whether there was a recent shark attack) have large and consistent effects on the same outcomes (e.g., voting behavior) overlooks the fact that if such effects were large, we should expect them to interact in complex ways. Hence, we should expect it to be very difficult to observe stable and replicable effects [196].

Besides consistency with observed data, choice of model representation should be consistent with prior knowledge. However, conventional approaches to estimating causal effects given an observational model are "memory-less" in the sense that prior estimates of an effect of interest, such as gained from past experiments in a similar vein, are not generally integrated into inference. Combined with incentives to publish surprising results [69, 173] and the inflated probability of observed effects to be overestimates in small sample size studies (Section 3.1), this can result in published effects that seem suspiciously big in light of prior domain knowledge.

Machine learning. *Optimizing for predictive accuracy in theory does not require well approximating a true DGP. However, researchers' persistent use of (asymptotic) theory for model motivation and holdout sets for estimating out-of-sample error lead to unrealistic beliefs about the predictability of real world processes. Perceived errors in learning also arise from failures to explicitly represent a priori human biases in the form of expectations about what constitutes a valid predictor for the task. There is a novel emphasis on ML research's convergence on foundation models that combine pre-trained representations with application-specific information but are unwieldy in their size and difficult to analyze and parameterize.*

The biggest point of contrast between representations in supervised learning in ML and social psychology is that the former traditionally do not assume that the learning process is "realizable" [170] in the sense that the true DGP is in the set of learnable functions (i.e., the hypothesis space), nor even that the fitted function appropriately approximates the true DGP. Instead the goal of learning is to search the hypothesis space for a function that minimizes error with respect to a concept class representing all the functions that behave exactly like the true function [197]. The standardization of learning in ML as developing the best predictor on a training data set and testing it on an unseen dataset assumed to be from the same distribution has motivated selecting model representations for their theoretical guarantees (e.g., on generalization ability like PAC learning, convergence, etc.), typically defined relative to worst-case bounds. Like psychologists' preoccupied with between-subjects designs, asymptotic statistical guarantees can be irrelevant outside of the idealized settings in which they are defined; e.g., a learner that's better asymptotically is worse under limited data [62]. Recently the limitations of classical theory for explaining the generalization performance of DNNs (e.g., [16, 17, 53, 119, 212]) are being acknowledged, leading to lines of theoretical work that explore different explanations.

One of the most commonly cited deficiencies attributed to model representations in applied ML is assuming a static relationship between the predictor variables and the outcome, which gives rise to conventions like shuffling input data to create training and test sets [8]. This assumption makes models vulnerable to concept drift [206] (a.k.a. covariate shift [28] or distribution or dataset shift [186]), where predictions are inaccurate post-hoc due to non-stationarity in the real-world relationship between the inputs and outputs due to temporal changes (e.g., [135]), behavioral reactions (e.g., [158]), or other unforeseen dynamics. Distribution shift can lead to poorly calibrated estimates of the uncertainty of model performance [154], similar to how choosing estimators by convention rather than guided by one's inference goal (see Section 3.3) biases uncertainty estimates for effects observed in psych experiments.

Distribution shift motivates increasing focus in research on how different models fare at out-of-distribution error and their robustness to adversarial manipulation, i.e., small changes to an input in feature space that dramatically change the predicted output [15, 42, 168, 190]. While these concerns can seem unique to ML as a more applied field than social psychology, recent results related to adversarial non-robustness [116], underspecification [52], shortcut learning [74], simplicity bias [178], and competency problems [72] suggest that beliefs about the true DGP in predictive modeling as in ML are not necessarily as distinct from causal attribution modeling as past comparative accounts (e.g., [36]) imply.

For example, one understanding of concept drift that we can relate to the so-called crud factor in psychology is that the concept of interest (or target task) in an ML pipeline for discriminative learning often depends on a complex combination of features that are not explicitly represented in the model. Geirhos et al. [74] use "shortcut learning" to refer to a tendency for ML models to learn simple decision rules (e.g., [9, 121, 138]) that perform well on standard benchmarks (based on the sorts of real correlations that exist between many variables in Meehl's conception). The problem is that singular predictive features mined in training data often do not perform as well in more challenging testing situations, where a human assumes that successful performance requires combinations of features (e.g., derived from several different object attributes in object recognition). Shortcut learning and related vulnerabilities to adversarial manipulation imply not a failure in learning from a modeling standpoint, nor even a failure of a fitted function to generalize [74], but a mismatch between a human's conception of critical, stable properties that predict under the true DGP and those that drive the predictions of the fitted model [52, 74, 116].

A related theory is underspecification [52]: specifically, a failure to represent in the learning pipeline which inductive biases are more desirable to constrain learning. Underspecification occurs when predictors with equivalent performance on i.i.d. data from the same distribution as training degrade non-uniformly in performance when probed along practically relevant dimensions [52]. Underspecification is distinct from forms of distribution shift that may give rise to shortcut learning, such as the presence spurious features in the training data that are not associated with the label in other settings. Instead, it captures how a single learning problem specification can support many near optimal solutions but which might have different properties along some human relevant dimensions like fairness or interpretability [169].

A common approach to overcoming poor generalization of a model out of distribution is to combine multiple model representations. Representation learning—automated, untrained learning of input representations (i.e., generic priors) on huge datasets that capture structure in domains like language or vision—reduces the difficulty of achieving high accuracy in domains where labeled data is costly [21]. "Fine-tuning" the pretrained models for domain-specific applications has become a standard practice based on the performance that can be achieved over conventional domain-specific learning pipelines [98, 188, 207]. Though highly diverse input data imbibes models adapted to these "foundation models" [31] with inductive biases that improve extrapolation, a challenge is that fine-tuning requires setting poorly understood parameters, making results hard to replicate [61]. For example, the robustness of a fine-tuned model has been found to vary considerably under small changes to hyperparameters [207]. Related to foundation models is a concern that the convergence in deep learning research around large DNN model architectures with minimal task-specific parameters [31] doubles down on an approach that imposes unreasonable environmental [19, 187] and research opportunity costs [19].

3.3 Model selection and evaluation

We describe concerns that arise in model-based inference, including explicit and implicit choices of objective function, optimization approach and evaluation metric.

Social psychology. *Bad claims in social psychology research can stem from the treatment of conventional approaches to model-based inference as a black box for consuming data and outputting inferences [11, 88, 128]. They may implicitly use statistical significance as a criterion for deciding what to report but ignore practical significance.*

"Inference by convention" can produce misleading claims without outright cheating or motivated reasoning. Researchers rarely motivate the estimators and loss functions they use for their specific inference goals. For example, conventional use of maximum likelihood estimators based on their consistency [203] may lead researchers to overlook critical assumptions required for these estimators to be well calibrated (i.e., have sampling distributions which are asymptotically normal). Relying on analytical approaches to optimization brings convenience, but relying on approaches where theory is formulated to describe their desirable properties pre-experimentally (i.e., before data is collected) can lead researchers to use procedures that do not achieve the desired characteristics on their particular data [22, 201].

A qualitatively different source of misleading claims is researchers' implicit use of statistical significance as a coarse objective function. This introduces associated misinterpretations and deficiencies of p-values (e.g., [76, 96, 112, 113, 200]), and detracts attention from questions of practical significance. The use of p-values and "statistical significance" in psychology research is described as fundamentally confused in that rejection of straw-man null hypotheses is inappropriately taken as evidence in favor of researchers' preferred alternatives [91, 113, 128]. Hypothesis testing is sometimes used as a sort of truth mill in psychology [80, 82].

Related problems include failing to acknowledge that as a random variable, p can vary considerably even under idealized replication [32, 85, 91, 145, 177], such that the difference between significant and not significant is not itself significant. Researchers also overlook the fact that for p to be a valid estimate of the probability of observing an effect as large or larger than that seen, all assumptions about the test and observational process must hold [4, 97, 160].

The well-known multiple comparisons problem, in which researchers neglect to control for the number of tests they run, reduces the validity of p values whenever researchers are engaging in data-dependent exploration of multiple analysis paths that involve statistical testing [81]. Implicit optimization for significance in which researchers are essentially searching through a garden of forking paths for analysis specifications that achieve significance as a sort of quasi-optimization approach [83] means that conventional interpretations of fitted models and statistical tests on parameter estimates, which assume data are only used once, will not hold. At the highest level, bias affects the published record when researchers decide whether to report results based on whether effects of interest are significant (publication bias or the "file drawer effect") [173].

Machine learning. *In ML, claims depend heavily on explicit choices about inference approach due to the strong influence of hyperparameters, initial conditions, and other configuration details on performance in non-convex optimization. However, researchers also may exploit flexibility in designing performance comparisons in order to achieve superior performance for their contributed approach relative to alternatives [107, 176].*

In contrast to loss functions in simple regression models, ML models tend to have high dimensional non-convex loss. While this does not prevent them from generalizing [46], it can prevent the "plug-and-play" approach to inference that arises in psychology. Optimizers—algorithms that prescribe how to update parameter values like weights during inference to reduce the value of the objective on the training data—are critical to the accuracy gains seen in recent years. However, finding good local minima to make non-convex optimization tractable requires setting various opaque hyperparameters that influence how the loss landscape is traversed, and there are generally no guarantees that a local solution identified under a certain computational budget is close to a global optimum.

For an optimization approach such as stochastic gradient descent (SGD), hyperparameters such as the learning rate affect how quickly it learns the local optima of a function: too high a rate means the function cannot converge, too low a rate it may require too long to converge [34]. Optimizer performance is sensitive to such hyperparameter settings and the initial state from which a search of the loss landscape proceeds [45, 175]. Adaptive optimizers (e.g., Adagrad, Adam) allow critical hyperparameters like learning rate to vary for each training parameter, inducing a new dynamical system with each run that makes it difficult to assess what aspects of an inference pipeline lead to better performance results.

For example, hyperparameter tuning is a computationally expensive task [194], hence it induces uncertainty about how much a solution might differ under a larger computational budget or different parameter settings. This makes fixing the computational budget for all modeling steps, including hyperparameter tuning, an important step to enabling a fair comparison between optimizers. Some recent work finds that given a fixed computational budget, choosing the best optimizer for a task with the default parameters performs about as well as choosing any widely used optimizer and tuning its hyperparameters, questioning claims of state-of-the-art performance of newly introduced optimizers across tasks [175]. Similarly, sufficient hyperparameter optimization can mostly eliminate claimed performance differences in GANs [134], and better hyperparameter tuning on baseline implementations can eliminate evidence of performance advantages of new learning methods [103, 141]. Liao et al. [130] use the broader term "implementation variation" to refer to how variations in how inference techniques are implemented—including use of specific software frameworks and libraries, metric scores, and implementation "tricks" [5, 103]—can affect their performance in evaluations. A related concern in subareas like reinforcement learning is when researchers overlook sources of inherent stochasticity in the training process and evaluation environment [125, 146, 205].

Other optimization and inference issues concern the external validity of the functions that are learned: will they predict well on unseen data? In the absence of a theoretical foundation for understanding DNN performance, exploratory empirical research is used to identify proxies for properties like learnability and generalizability (e.g., [119, 212]). Recent results show how counter to classical expectations about overfitting, minimizing training error without implicit regularization over overparameterized models tends to result in good generalization despite the empirical optimization problem being underdetermined and the presence of many global minima (e.g., [183]). Entanglements between optimization methods and statistical properties of the solutions they find drive a new theoretical agenda aimed at explaining how implicit biases induced by optimization algorithms contribute to performance [149, 183, 212]. Some findings about emergent properties have been critical; for example, related to shortcut learning (Section 3.2), SGD has been shown to exhibit "simplicity bias"—a preference for learning simple predictors first which results in neural nets exclusively relying on the simplest features, for example, the color and texture of images, and remaining invariant to all predictive complex features, for example, the shape of objects in images [122, 178].

Other concerns with external validity arise whenever an explicitly chosen objective function may not actually be a good proxy for the metric of interest in the use of the models. For example, cross-entropy loss is often used as a loss function, whereas the evaluation metric of interest is often classification error or AUCPR. This is referred to as "loss-metric misalignment" and can hurt the generalization of ML models [111]. More generally, the conventional focus in reporting single scalar error measures risks overlooks important error variation (e.g., [67]).

Internal and external validity issues can arise from leakage–broadly, the use of information from the test data in training–paralleling the reuse of data for choosing a model and evaluating its fit in psychology. Leakage and related issues can arise through inappropriate use of cross-validation (CV), commonly used to refine and evaluate the performance of ML models. CV involves partitioning a dataset into multiple "folds" and iteratively using one fold as the test set and the others as training sets, then averaging the performance evaluations obtained for an overall measure of performance. However, CV underestimates error on test data results when a single CV procedure is used for model tuning and estimating error at once [43, 102]. Failure to carefully consider which steps involved in model training should also be performed on each fold during CV can bias error estimates on unseen test data [102], as can contaminating the procedure with future data in time series applications [23]. More generally, the use of CV for performance evaluation has been shown to lead to overoptimistic results in the presence of dependencies between the training and test set under certain conditions, captured by optimism [68, 137]: the expected amount by which training error will under-estimate true error (i.e., test error) obtained on data not used to train the model. The difference between these errors boils down to a term describing how the training data and its fitted values (i.e., predicted values on the training data) covary [137]. Not unlike how low power settings lead to overestimates of observed effects in psychology, under such dependencies, underestimating test error from CV becomes more likely.

Other issues occur in performance comparisons of models or algorithms. Similar to data issues in psychology, sampling error can be overlooked, including low power in performance comparisons [41] and failure to acknowledge that performance estimates on the standard train-test splits common in benchmark datasets may not hold for randomly created train-test splits [93].

Finally, implicit optimization for good performance results can also occur in ML. Improving performance on benchmark datasets, which have been widely adopted across multiple subfields of ML (e.g., Computer Vision [58], Natural Language Processing [202], Graph Machine Learning [110]) and been thought to have caused most major ML research breakthroughs in the last 50 years [63], is the de facto way that researchers showcase improvement in model performance to get published in top conferences and journals [161, 176]. This can create incentives for researchers to implicitly optimize inference around a goal of seeing their new technique rank best in performance in an evaluation, such as selectively reporting results to highlight the best accuracy achieved (see Section 4 below), choosing among performance measures conditional on results, or failing to acknowledge how simpler baselines perform relative to a new approach (e.g., how well the "language prior," the prior distribution over labels [95], performs in a popular visual question answering task [7]).

4 COMMUNICATION OF CLAIMS

Sources of error become problematic when not acknowledged due to conventions that suppress uncertainty and limit reproducibility.

Social psychology. *The salient contribution of a social psychology experiment can be thought of as a stylized fact: a statement that is presumed to be generally true about some aspect of the world, and replicable [79, 106]. Generic statements appear to be the norm in psychology articles according to one recent study [57], with communication deficiencies attributable to two sources: 1) authors failing to acknowledge exploration of multiple analysis paths contingent on the data, and 2) authors' tendency to downplay inherent dependencies and uncertainty when describing results.*

For example, because stylized facts derive from the results of experiments in laboratory-like environments, often on non-representative samples [79], credible reporting of results would emphasize that they may not generalize more broadly and document the specific conditions studied [182]. Instead, however, researchers routinely restate their findings in broad terms in writing an article, referring to how an intervention or specific trait affects "people" or entire groups (e.g., Whites) without hedging [57, 99, 167], and ignoring potential variation within groups.

Authors can perpetuate p-value fallacies when they write about effects as if present or absent (e.g., [27]) or overinterpret alternative hypotheses [192]. Or they may imply that a lack of significance is evidence of an absence of effect [3, 27] or that there is a significant difference between significant and non-significant results [85].

Finally, while sharing of data and analysis code has increased in psychology in recent years, many authors have not adopted such sharing (e.g., [193]). When authors don't publish data or analysis code they used to arrive at a conclusion, readers cannot as easily identify problems or replicate the work, potentially slowing the rate at which errors that invalidate claims are caught [78, 109, 144].

Machine learning. *Communication concerns in ML are largely analogous to those in SP, taking the form of tendencies to omit reporting of trial and error over the modeling pipeline and evaluation metrics (leading to biased claims about model performance) and to downplay dependencies and uncertainty affecting performance.*

Instead of p-values, ML researchers often report point estimates of performance without quantifying uncertainty [2, 130, 162]. Researchers also underreport on the conditions that gave rise to the point estimates, such as hyperparameter and computational budget settings in non-convex optimization. This can result in performance results for which the source of empirical gains is unclear or misattributed [132]. As examples, authors often do not report the number of models trained and the negative results found before the one they highlight is selected [2, 176]. Authors may cut corners since computing uncertainty and variance in ML models can incur significant computational costs, especially for large ML models [2, 34]. When not presented along with an estimate of the uncertainty of model performance arising from sources of variation like the choice of train-test split [93], the computational budget [60], the choice of hyperparameter values and the random initialization of ML models [45, 134, 175], point estimates of performance represent the best-case rather than expected model performance. Worse, researchers sometimes apply CV to tune a model then report the best performing model's error on the training set (i.e., the "apparent error") as if it were cross-validated error [148].

Similar to psychology, researchers may be tempted to speculate about causes without couching them in speculative terms [132]. Overgeneralization occurs from the loose connection between a task (e.g., reading comprehension, image classification) given in colloquial and anthropomorphic terms as what a model has learned to do, and much more specific definition of the learning problem [130, 132] for which publishable results were achieved. For example, using "reading comprehension" to refer to what a model has attained is misleading when the model may not have used what a human would call critical information, like the text it is "comprehending" [123] (see Section 3.3). More broadly, claims about model performance are rarely evaluated in the context of the real-world applications they are implied to generalize to [130].

There are new and analogous issues to the lack of open data and code in social psychology [66, 93, 100, 171, 185]. Details about dataset limitations that can threaten external validity [19, 73, 156] are often unreported in ML literature (Section 3.1), perhaps because new techniques for model creation are valued over documenting datasets [174]. But closer to data sharing issues in psychology, checking computational reproducibility of results requires making the complete code and data available with published papers [38]. Recent work aims to improve reproducibility using reproducibility checklists, documentation checklists, community challenges, and workshops [73, 143, 159]. While assessing replication in the social sciences is not trivial (e.g., [184]), a somewhat unique challenge in ML is that with the creation and widespread use of large ML models requiring significant computational resources [19], especially in NLP tasks, it becomes impossible for many researchers to attempt replicating certain results.

5 SUMMARY AND DISCUSSION

While others have drawn analogies between reproducibility in ML and social sciences (e.g., [105, 107, 210]), by drawing connections at various decision points in the learning process our work is unique in scope and aims. Table 1 summarizes results, including a number of concerns related to human-mediated error in the learning pipeline that arise for similar reasons across social psychology and ML.

We see evidence of different, but analogous ways in which researchers place undue confidence in particular statistical methods. In ML, the use of a train/test split and cross validation can give the illusion that the inherent inability to know performance on unseen data is manageable. In social psych, belief in the power of randomized sampling and statistical testing leads researchers to overlook the importance of satisfying other assumptions, like sampling stimuli. Motivating choices like model representation using asymptotic theory without considering its applicability to the specific inference problem is conventional. In both cases, researchers' trust in methods is undergirded by unrealistic expectations about the predictability of real-world behavior and other phenomena. Social psychologists ignore the "crud factor" [139] and improbability that multiple predictors thought to have large effects on the same outcome would not also correlate with one another [196]. ML researchers seem to embrace the crud factor by recognizing the importance of using many predictors to avoid overfitting when the signal from any one predictor is likely to be small [53], but have been slow to part with i.i.d. assumptions.

In addition, norms around what is publishable in each field incentivize researchers to hack results to meet implicit objectives like statistical significance of hypotheses or better than SOTA performance, to the detriment of practical significance or external validity. Important dependencies in the analysis process–from types of data filtering and reuse to unacknowledged computational budgets or unspecified populations–are often overlooked, such that results cannot be taken at face value at worst and do not generalize as assumed at best. Overgeneralization and suppression of uncertainty via binary statements about the presence of effects or rank of model performance relative to baselines are common in reporting results.

One reason these results are useful is because as researchers move toward integrative modeling, with psychologists adopting checks of predictive accuracy and ML researchers adopting statistical design, testing, and measurement modeling, it will be important that they also grasp limitations and common blindspots associated with the new methods they adopt. The value of more interdisciplinary methods use is threatened if researchers assume that using integrative approaches must increase rigor, because "now we do statistical testing," "now we do human subjects experiments," or "now we use a test/train split."

In the more immediate term, our results can be used to identify points where solutions might be common despite differences in the two fields. On a deeper, structural level, our results can also be used to develop conceptual frameworks for describing issues in learning from data and trying to propose reforms. We consider both below.

5.1 Opportunities for common solutions

There are various concrete adjustments to conventional practice that are well motivated in both psychology and machine learning:

- *Pre-registration.* Pre-registration has gained widespread visibility in psychology, where it is designed to limit unreported data dependent decisions in analysis. Developing a similar approach for ML could curb threats to internal validity like selecting performance measures to report or data filtering procedures based on how the end results compare to SOTA.
- *Reporting templates for learning dependencies.* Reporting templates for datasets and models (e.g., [73, 143]) have been proposed in ML, but similar reporting standardizations, especially if required by large ML conference, could be used to addressed failures to report computational budgets, initial conditions, hyperparameter values and other dependencies. In psychology, beyond defining the population for generalization [182], standardized reporting of stimuli sampling assumptions and selection of subject sample sizes when NHST is used are well motivated.
- *Alternatives to reporting p-values.* Strategies for communicating uncertainty in results without using p-values, such as estimation [51] or reframings of p-values designed to emphasize their conditionality and information content [160], are likely to be just as helpful in ML especially given suggestions (and concerns) that NHST will become more widely used [26, 67].

In addition, though it will take different forms, better training on how to validate that a given learning problem is well approximated by asymptotics, and the importance of hedging claims when validation is not feasible, is well motivated in both fields, as are continued efforts to encourage sharing of data and all code artifacts upon submission.

5.2 Irrefutable claims, latent expectations

On a deeper level, claims researchers are making in both psych appear to be irrefutable both by design and convention. In social psychology, this manifests as papers that set out to confirm hypotheses that associations will exist, or be in a certain direction, rather than mechanistic accounts that enable more specific predictions. When hypotheses provide only weak constraints on researchers' ability to find confirming evidence *and* there is flexibility in the analysis process (not to mention pressures prioritizing counterintuitive effects and publishing positive evidence [69, 173]), "false positive psychology" [180] is not a surprising result. Consider how much more difficult, and even impossible, it for those who wish to refute, rather than support, a given theory: showing no association, for example, means providing evidence for a point prediction of null effect. At the same time, in the absence of well motivated stimuli sampling strategies, well-defined target populations, and attempts to model other sources of contextual variation, assuming that claims made about any particular set of parameter estimates obtained through analyzing experiment results generalize beyond that particular set of participants, stimuli, etc. is not credible.

Turning to ML, the cultural root cause of many of the pitfalls and reproducibility failures seems to derive from a similar tolerance for irrefutable contributions, manifesting in ML as a confusion between engineering artifacts and scientific knowledge. Consider a typical supervised ML paper that shows that an innovative algorithm, architecture, or model achieves a certain accuracy on a benchmark dataset. Even if we assume the reported accuracy is optimistic for the various reasons discussed above, the researcher has contributed an engineering artifact, a tool that the practicing engineer can carry in their toolbox based on its superior performance to the state-of-the-art on a particular learning problem. New observations based on additional data cannot refute the performance claim of the given algorithm on the given dataset, because the population from which benchmark datasets are drawn are rarely specified to the detail needed for another sample to be drawn [120]. Attempts to collect a different sample from an implied population to refute claims are rare; when they have been attempted, researchers have found that the original claims no longer hold on the new samples [165]. Further, when researchers have tried to compare model performance across benchmark datasets, they have found that results on one benchmark rarely generalize to another, and can be extremely fragile [56, 195].

At a higher level, analogies between human and artificial intelligence are embedded in AI and ML culture, but without theory that can render them refutable. Researchers rationalize post-hoc that ML approaches capture critical aspects of human consciousness (e.g., [94]) and describe new algorithms as inspired by human cognition (e.g., [20]). Without a priori specification of the neuro-computational processing involved in high level cognition [166], whether algorithms intended to instantiate human-like mechanisms succeed in a human-like way is entirely speculative.

The acceptance of non-refutable research claims as research contributions, as in social psychology, creates a culture in which other methodological issues amplify the difficulty of building generalizable knowledge. Hubris from beliefs that big data renders modeling requirements like uncertainty quantification unnecessary [33, 56, 131], a lack of rigor in evaluation [130, 176], and over-reliance on theory [89, 132] may leave ML plagued with reproducibility and generalization issues. One potential bright spot lies in widespread recognition that the field is lacking foundational statistical theory to explain DNN performance. Awareness of the shaky theoretical ground could naturally encourage a more cautious mindset among researchers, but only if pressures to make bold claims from structural incentives that encourage "planting one's flag" before others do [176] don't outweigh the trend toward embracing uncertainty.

We propose that characterizing the conventions that give rise to irrefutable claims as forms of underspecification can help point researchers in both fields toward new methods to address what is missing. By underspecification, we mean that some aspect of the learning problem has not been formalized to an extent that allows it to be solved. In particular, the role of human expectations in defining "success" in learning is typically ignored.

Consider how colloquially ML has been framed as freeing researchers from theorizing how well a fitted function may capture critical structure in the true DGP. Many of the challenges being identified by critics suggest that the reality of non i.i.d. test data appears to be pushing ML researchers in "purely" predictive areas back towards philosophies underlying causal statistical inference. Recent definitions of underspecification [52], shortcut learning [74], and adversarial vulnerability [116] motivate the need to elicit and impose constraints on what is learned. The most natural source of these constraints is the human who will assess and interpret the results of learning. This means dissatisfaction with the functions ML models are learning is a phenomena that cannot be blamed on the methods themselves: it indicates a mismatch between human expectations of the learning constraints and the actual hypothesis space and optimization process.

In social psychology, DGPs are modeled, if only as a symptom of using conventional inference. However, we see instead a displacement of prior knowledge when designing and interpreting experiments, where a priori expectations about how big an effect could be are often overlooked, and a failure to acknowledge how the styles of research being rewarded in the field, such as showing that many small interventions can have large effects on a class of outcomes, are incompatible with common sense expectations of correlated effects. A relevant lesson applicable to both fields is that human agency cannot be abstracted away to achieve good inference, and that essential human inductive biases and domain knowledge remain unintegrated in the state-of-the-art. There is little reason to believe that taking steps toward integrative modeling will greatly improve practice without also addressing the role of the researcher's own beliefs in making sense of learning. In both fields, research around how to elicit human expectations and bring them to bear on a learning process more systematically are well motivated.

5.3 Epistemological gaps and rhetorical risks

Ideas of a universal method of statistical inference [88] (implying a "bad methods" explanation for errors) or that methods for justifying claims against evidence are "objectively correct" if we remove or restrict the human element [37, 64] (implying a "bad humans" explanation) arise in light of the inherent uncertainty faced in producing scientific claims. It is natural for fields to amass signals thought to be proxies of trustworthiness to enable judging work at the time of publication, when how well a claim replicates or generalizes is not known. However, a fundamental challenge in doing this is the need for reformers to recognize the incompleteness of their own knowledge, so as to avoid overcertainty in attempts toward reform.

Consider how irrefutable theories and claims induce greater dependence in the fields on imperfect ways of validating claims. In social psychology, replication is an indirect test for whether effects persist under the same or similar conditions. However, experts do not necessarily agree on what constitutes successful replication [157], and intuitions can be proven wrong. For example, under a formal definition of a study's reproduciblity rate, reproducing experimental results doesn't necessarily indicate a "true" effect and lack of reproducibility does not necessarily indicate a "false" effect [59]. In ML, tests are similarly indirect but the stakes often higher: when an approach fails to perform as well as expected in the world, researchers may scrutinize the original claims, but at the expense of those affected in deployment and without a theoretical foundation for making precise statements about how the real world setting was surprising.

There is a need to accurately diagnose the fundamental problems, rather than addressing symptoms only, and avoid the sort of part-for-whole substitution in reforms that drive methodological overconfidence. As fields work toward consensus views on errors, uncertainty must be embraced. For example, debates over what core problem(s) preregistration addresses point to the challenge of determining when a given reform should have privileged status over others [147, 181, 191]. In ML, the value of benchmarks is similarly complex. While blamed for many issues with external validity (e.g., [161]), there is evidence that performance on benchmarks remains a stable predictor of certain forms of generalization [163, 164, 208]. With foundational models benchmarks may naturally become higher quality and more diverse [31].

Trusting a method (whether it be a statistical idea such as Bayesian inference or causal identification, or an ML idea such as deep learning or cross-validation) without examining the applied context can mislead researchers by implying that better learning from data can be achieved in singular or simple ways. It can be that more careful researchers tend to use more sophisticated methods, which will show up as a correlation between methodological sophistication and the quality of research—but this unfortunately can also create an opening for methods to be used as a signal of research quality even when that is not the case. For example, it makes sense for open-science reforms to be supported by researchers who do stronger work (and there is evidence from betting markets that experts can predict reproducibility with some accuracy [65]) and opposed by those whose work has failed to replicate (for example, [211]), which would lead to open-science practices themselves being a marker of research quality. On the other hand, honesty and transparency are not enough [78]: all the openness and preregistration in the world won't endow replicability to a psychology study with a high ratio of noise to signal, which can happen with experiments whose designs focus on procedural issues (e.g., randomization), to the detriment of theory and measurement. Open-science practices can be a signal of replicability without that holding in the future.

Beyond greater awareness of rhetorical risks in reform, concrete steps can be taken. For example, Devezer et al. [59] propose that colloquial statements about problems and solutions be accompanied with formal problem statements and results, and provide a template of questions to guide researchers in doing so. Directing more effort toward rigor in reform arguments stands to enable quicker identification of logical errors, misintepretations of statistical constructs, or other forms of fallibility in attempts to steer a field back on track. Both ML and social psychology could benefit.

6 CONCLUSION

Our analysis of learning errors across psychology and supervised machine learning points to fundamental blindspots in inductive learning related to overtrusting theory and conventional practice for producing (irrefutible) claims and simplified assumptions of real world variation, among others. We argue that many learning errors are errors at the method-human interface, especially underspecification of human biases, which integrative approaches alone cannot solve without greater awareness of these blindspots.

7 ACKNOWLEDGMENTS

We thank Jake Hofman and Cyril Zhang for comments on a draft. Hullman's work is supported by the National Science Foundation under award IIS-1930642 and by a Microsoft Faculty Fellowship. Narayanan's work is supported by the National Science Foundation under award IIS-1763642. Gelman's work is supported by the Office of Naval Research.

REFERENCES

[1] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 801–809.

[2] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems* 34 (2021).

[3] Douglas G Altman and J Martin Bland. 1995. Statistics notes: Absence of evidence is not evidence of absence. *Bmj* 311, 7003 (1995), 485.

[4] Valentin Amrhein, David Trafimow, and Sander Greenland. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *American Statistician* 73, sup1 (2019), 262–270.

[5] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. 2020. What matters for on-policy deep actor-critic methods? a large-scale study. In *International conference on learning representations*.

[6] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics*. Princeton university press.

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[8] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[9] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International Conference on Machine Learning*. PMLR, 233–242.

[10] Linda Babcock and George Loewenstein. 1997. Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives* 11, 1 (1997), 109–126.

[11] David Bakan. 1966. The test of significance in psychological research. *Psychological bulletin* 66, 6 (1966), 423.

[12] John A Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology* 71, 2 (1996), 230.

[13] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[14] Roy F Baumeister, Kathleen D Vohs, and David C Funder. 2007. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science* 2, 4 (2007), 396–403.

[15] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in Terra Incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 456–473.

[16] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854.

[17] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. 2018. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems* 31 (2018).

[18] Samuel J Bell and Onno P Kampman. 2021. Perspectives on Machine Learning from Psychology's Reproducibility Crisis. *arXiv preprint arXiv:2104.08878* (2021).

[19] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[20] Yoshua Bengio. 2017. The consciousness prior. *arXiv preprint arXiv:1709.08568* (2017).

[21] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[22] James O Berger and Robert L Wolpert. 1988. The Likelihood Principle. IMS.

[23] Christoph Bergmeir and José M Benítez. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* 191 (2012), 192–213.

[24] José M Bernardo and Adrian FM Smith. 2001. Bayesian Theory.

[25] Ryan Bernstein. 2021. Drawing maps of model space with modular Stan. (2021). https://statmodeling.stat.columbia.edu/2021/11/19/drawing-maps-of-model-space-with-modular-stan/

[26] Daniel Berrar and Werner Dubitzky. 2019. Should significance testing be abandoned in machine learning? *International Journal of Data Science and Analytics* 7, 4 (2019), 247–257.

[27] Lonni Besançon and Pierre Dragicevic. 2019. The continued prevalence of dichotomous inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[28] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research* 10, 9 (2009).

[29] María J Blanca, Rafael Alarcón, and Roser Bono. 2018. Current practices in data analysis procedures in psychology: What has changed? *Frontiers in psychology* 9 (2018), 2558.

[30] Niall Bolger, Katherine S Zee, Maya Rossignac-Milon, and Ran R Hassin. 2019. Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General* 148, 4 (2019), 601.

[31] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[32] Dennis D Boos and Leonard A Stefanski. 2011. P-value precision and reproducibility. *American Statistician* 65, 4 (2011), 213–221.

[33] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. 2021. Accounting for variance in machine learning Bbenchmarks. In *Machine Learning and Systems (MLSys)*.

[34] Xavier Bouthillier and Gaël Varoquaux. 2020. *Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020*. Ph. D. Dissertation. Inria Saclay Ile de France.

[35] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[36] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (2001), 199–231.

[37] Matthew J Brown. 2013. Values in science beyond underdetermination and inductive risk. *Philosophy of Science* 80, 5 (2013), 829–839.

[38] Jonathan B Buckheit and David L Donoho. 1995. Wavelab and reproducible research. In *Wavelets and statistics*. Springer, 55–81.

[39] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[40] Katherine S Button, John Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* 14, 5 (2013), 365–376.

[41] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595* (2020).

[42] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

[43] Gavin C Cawley and Nicola L C Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11 (2010), 2079–2107.

[44] Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46, 1 (2014), 112–130.

[45] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. 2019. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446* (2019).

[46] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. 2015. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*. PMLR, 192–204.

[47] Herbert H Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior* 12, 4 (1973), 335–359.

[48] Jacob Cohen. 1992. Statistical power analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101.

[49] Jeremy R Coyle, Nima S Hejazi, Ivana Malenica, Rachael V Phillips, Benjamin F Arnold, Andrew Mertens, Jade Benjamin-Chung, Weixin Cai, Sonali Dayal, John M Colford Jr, et al. 2020. Targeting learning: robust statistics for reproducible research. *arXiv preprint arXiv:2006.07333* (2020).

[50] Kate Crawford. 2017. The trouble with bias. (2017). https://www.youtube.com/watch?v=fMym_BKWQzk NIPS 2017.

[51] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[52] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[53] Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. 2021. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355* (2021).

[54] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.

[55] Erica Dawson, Thomas Gilovich, and Dennis T Regan. 2002. Motivated Reasoning and Performance on the was on Selection Task. *Personality and Social Psychology Bulletin* 28, 10 (2002), 1379–1387.

[56] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *arXiv preprint arXiv:2107.07002* (2021).

[57] Jasmine M DeJesus, Maureen A Callanan, Graciela Solis, and Susan A Gelman. 2019. Generic language in scientific communication. *Proceedings of the National Academy of Sciences* 116, 37 (2019), 18370–18377.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 248–255.

[59] Berna Devezer, Danielle J Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. 2020. The case for formal methodology in scientific reform. *Royal Society Open Science* 8, 3 (2020), 200805.

[60] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004* (2019).

[61] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305* (2020).

[62] Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 10 (2012), 78–87.

[63] David Donoho. 2018. 50 years of data science. (2018).

[64] Heather Douglas. 2016. Values in science. In *The Oxford Handbook of Philosophy of Science*.

[65] Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112 (2015), 15343–15347.

[66] Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics* 5 (2017), 471–486.

[67] Chris Drummond. 2006. Machine learning as an experimental science (revisited). In *AAAI Workshop on Evaluation Methods for Machine Learning*. 1–5.

[68] Bradley Efron. 2004. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* 99, 467 (2004), 619–632.

[69] Daniele Fanelli. 2010. "Positive" results increase down the hierarchy of the sciences. *PloS one* 5, 4 (2010), e10068.

[70] Gregory Francis. 2012. The psychology of replication and replication in psychology. *Perspectives on Psychological Science* 7 (2012), 585–594.

[71] Gregory Francis. 2012. Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review* 19 (2012), 975–991.

[72] Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency problems: On finding and removing artifacts in language data. *arXiv preprint arXiv:2104.08646* (2021).

[73] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[74] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[75] Andrew Gelman. 2012. Ethics and statistics: Ethics and the statistical use of prior information. *Chance* 25, 4 (2012), 52–54.

[76] Andrew Gelman. 2013. P values and statistical practice. *Epidemiology* 24, 1 (2013), 69–72.

[77] Andrew Gelman. 2015. The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* 41, 2 (2015), 632–643.

[78] Andrew Gelman. 2017. Honesty and transparency are not enough. *Chance* 39 (2017), 37–39. Issue 1.

[79] Andrew Gelman. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* 44, 1 (2018), 16–23.

[80] Andrew Gelman and John B. Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651.

[81] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 348 (2013).

[82] Andrew Gelman and Eric Loken. 2014. Ethics and statistics: The AAA tranche of subprime science. *Chance* 27, 1 (2014), 51–56.

[83] Andrew Gelman and Eric Loken. 2014. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist* 102, 6 (2014), 460.

[84] Andrew Gelman, Daniel Simpson, and Michael Betancourt. 2017. The prior can often only be understood in the context of the likelihood. *Entropy* 19 (2017), 555.

[85] Andrew Gelman and Hal Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician* 60, 4 (2006), 328–331.

[86] Andrew Gelman and David Weakliem. 2009. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist* 97, 4 (2009), 310–316.

[87] Gerd Gigerenzer. 2022. We need to think more about how we conduct research. *Behavioral and Brain Sciences* 45 (2022).

[88] Gerd Gigerenzer and Julian N Marewski. 2015. Surrogate science: The idol of a universal method for scientific inference. *Journal of management* 41, 2 (2015), 421–440.

[89] Tom Goldstein. 2022. My recent talk at the NSF town hall focused on the history of the AI winters, how the ML community became "anti-science," and whether the rejection of science will cause a winter for ML theory. I'll summarize these issues below... http://archive.today/ryryU

[90] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[91] Steven N Goodman. 1993. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137, 5 (1993), 485–496.

[92] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[93] Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2786–2791.

[94] Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091* (2020).

[95] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[96] Sander Greenland. 2019. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *American Statistician* 73, sup1 (2019), 106–114.

[97] Sander Greenland and Zad Rafi. 2019. To aid scientific inference, emphasize unconditional descriptions of statistics. *arXiv preprint arXiv:1909.08583* (2019).

[98] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020).

[99] Kris D Gutiérrez and Barbara Rogoff. 2003. Cultural ways of learning: Individual traits or repertoires of practice. *Educational researcher* 32, 5 (2003), 19–25.

[100] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16.

[101] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems* 24, 2 (2009), 8–12.

[102] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[103] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[104] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences* 33, 2-3 (2010), 61–83.

[105] Mireille Hildebrandt. 2018. Preregistration of machine learning research design. Against P-hacking. (2018).

[106] Daniel Hirschman. 2016. Stylized facts in the social sciences. *Sociological Science* 3 (2016), 604–626.

[107] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.

[108] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. 2021. Integrating explanation and prediction in computational social science. *Nature* 595, 7866 (2021), 181–188.

[109] Bobby Lee Houtkoop, Chris Chambers, Malcolm Macleod, Dorothy VM Bishop, Thomas E Nichols, and Eric-Jan Wagenmakers. 2018. Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018), 70–85.

[110] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2021. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv:2005.00687 [cs, stat]* (Feb. 2021). http://arxiv.org/abs/2005.00687 arXiv: 2005.00687.

[111] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. 2019. Addressing the loss-metric mismatch with adaptive loss alignment. In *International conference on machine learning*. PMLR, 2891–2900.

[112] Raymond Hubbard and MJ Bayarri. 2003. P values are not error probabilities. *Institute of Statistics and Decision Sciences, Working Paper* 03-26 (2003), 27708–0251.

[113] Raymond Hubbard and María Jesús Bayarri. 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician* 57, 3 (2003), 171–178.

[114] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[115] Matthew Hutson. 2018. Has artificial intelligence become alchemy?

[116] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems* 32 (2019).

[117] John P. A. Ioannidis. 2008. Why most discovered true associations are inflated. *Epidemiology* 19 (2008), 640–648.

[118] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[119] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2019. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178* (2019).

[120] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.

[121] Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561* (2017).

[122] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. 2019. SGD on neural networks learns functions of increasing complexity. *Advances in neural information processing systems* 32 (2019).

[123] Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926* (2018).

[124] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 3819–3828.

[125] Khimya Khetarpal, Zafarali Ahmed, Andre Cianflone, Riashat Islam, and Joelle Pineau. 2018. Re-evaluate: Reproducibility in evaluating reinforcement learning algorithms. (2018).

[126] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[127] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[128] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

[129] Thomas Liao, Benjamin Recht, and Ludwig Schmidt. 2020. In a forward direction: Analyzing distribution shifts in machine translation test sets over time. (2020).

[130] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[131] Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2021. Significant improvements over the state of the art? A case study of the MS MARCO Document Ranking Leaderboard. (Feb. 2021). https://arxiv.org/abs/2102.12887v1

[132] Zachary C Lipton and Jacob Steinhardt. 2019. Research for practice: troubling trends in machine-learning scholarship. *Commun. ACM* 62, 6 (2019), 45–53.

[133] Eric Loken and Andrew Gelman. 2017. Measurement error and the replication crisis. *Science* 355, 6325 (2017), 584–585.

[134] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are gans created equal? a large-scale study. *Advances in neural information processing systems* 31 (2018).

[135] Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2021. Time Waits for No One! Analysis and Challenges of Temporal Misalignment. *arXiv preprint arXiv:2111.07408* (2021).

[136] John G Lynch Jr. 1982. On the external validity of experiments in consumer research. *Journal of consumer Research* 9, 3 (1982), 225–239.

[137] Momin M Malik. 2020. A hierarchy of limitations in machine learning. *arXiv preprint arXiv:2002.05193* (2020).

[138] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007* (2019).

[139] Paul E Meehl. 1967. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34, 2 (1967), 103–115.

[140] Paul E Meehl. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological reports* 66, 1 (1990), 195–244.

[141] Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589* (2017).

[142] Xiao-Li Meng. 2018. STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I) LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION. *The Annals of Applied Statistics* 12, 2 (2018), 685–726.

[143] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[144] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 1–9.

[145] Duncan J Murdoch, Yu-Ling Tsai, and James Adcock. 2008. P-values are random variables. *American Statistician* 62, 3 (2008), 242–245.

[146] Prabhat Nagarajan, Garrett Warnell, and Peter Stone. 2018. Deterministic implementations for reproducibility in deep reinforcement learning. *arXiv preprint arXiv:1809.05676* (2018).

[147] Danielle Navarro. 2020. Paths in strange spaces: A comment on preregistration. (2020).

[148] Marcel Neunhoeffer and Sebastian Sternberg. 2019. How cross-validation can go wrong and what to do about it. *Political Analysis* 27, 1 (2019), 101–106.

[149] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614* (2014).

[150] Matthew P Normand. 2016. Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology* 7 (2016), 934.

[151] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[152] Brian A. Nosek et al. 2015. Estimating the reproducibility of psychological science. *Science* 349 (2015), aac4716.

[153] Amy Orben and Daniël Lakens. 2020. Crud (re) defined. *Advances in methods and practices in psychological science* 3, 2 (2020), 238–247.

[154] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems* 32 (2019).

[155] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).

[156] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[157] Samuel Pawel and Leonhard Held. 2020. The sceptical Bayes factor for the assessment of replication success. *arXiv preprint arXiv:2009.01520* (2020).

[158] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[159] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research* 22 (2021).

[160] Zad Rafi and Sander Greenland. 2020. Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC medical research methodology* 20, 1 (2020), 1–13.

[161] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[162] Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808* (2018).

[163] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*

(2018).

[164] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International Conference on Machine Learning*. PMLR, 5389–5400.

[165] B Recht, R Roelofs, L Schmidt, and V Shankar. 2019. Unbiased look at dataset bias. ICML.

[166] James A Reggia, Garrett E Katz, and Gregory P Davis. 2020. Artificial conscious intelligence. *Journal of Artificial Intelligence and Consciousness* 7, 01 (2020), 95–107.

[167] Barbara Rogoff et al. 2003. *The cultural nature of human development*. Oxford university press.

[168] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. 2018. The elephant in the room. *arXiv preprint arXiv:1808.03305* (2018).

[169] Andrew Ross, Isaac Lage, and Finale Doshi-Velez. 2017. The neural lasso: Local linear sparsity for interpretable explanations. In *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*, Vol. 4.

[170] Stuart J Russell and Peter Norvig. 2003. Artificial Intelligence A Modern Approach. (2003).

[171] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS computational biology* 9, 10 (2013), e1003285.

[172] Kai Sassenberg and Lara Ditrich. 2019. Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science* 2, 2 (2019), 107–114.

[173] Jeffrey D Scargle. 1999. Publication bias (the" file-drawer problem") in scientific inference. *arXiv preprint physics/9909033* (1999).

[174] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[175] Robin M Schmidt, Frank Schneider, and Philipp Hennig. 2021. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*. PMLR, 9367–9376.

[176] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. (2018).

[177] S Senn. 2001. Two cheers for P-values? *Journal of Epidemiology and Biostatistics* 6, 2 (2001), 193–204.

[178] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 9573–9585.

[179] Galit Shmueli. 2010. To explain or to predict? *Statist. Sci.* 25, 3 (2010), 289–310.

[180] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, 11 (2011), 1359–1366.

[181] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2021. Pre-registration is a game changer. But, like random assignment, it is neither necessary nor sufficient for credible science. *Journal of Consumer Psychology* 31, 1 (2021), 177–180.

[182] Daniel J Simons, Yuichi Shoda, and D Stephen Lindsay. 2017. Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science* 12, 6 (2017), 1123–1128.

[183] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* 19, 1 (2018), 2822–2878.

[184] Peter M Steiner, Vivian C Wong, and Kylie Anglin. 2019. A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie* (2019).

[185] Victoria Stodden and Sheila Miguez. 2014. Provisioning Reproducible Computational Science. (2014).

[186] Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning* 30 (2009), 3–28.

[187] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13693–13696.

[188] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 843–852.

[189] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[190] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[191] Aba Szollosi, David Kellen, Danielle Navarro, Richard Shiffrin, Iris van Rooij, Trisha Van Zandt, and Chris Donkin. 2019. Is preregistration worthwhile? (2019).

[192] Denes Szucs and John Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience* 11 (2017), 390.

[193] Leho Tedersoo, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, et al. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data* 8, 1 (2021), 1–11.

[194] Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. 2019. Optimizer Benchmarking Needs to Account for Hyperparameter Tuning. *arXiv e-prints* (2019), arXiv–1910.

[195] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. IEEE, 1521–1528.

[196] Christopher Tosh, Philip Greengard, Ben Goodrich, Andrew Gelman, Aki Vehtari, and Daniel Hsu. 2021. The piranha problem: Large effects swimming in a small pond. *arXiv preprint arXiv:2105.13445* (2021).

[197] Leslie G Valiant. 1984. A theory of the learnable. *Commun. ACM* 27, 11 (1984), 1134–1142.

[198] Tyler J VanderWeele and Miguel A Hernán. 2012. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology* 175, 12 (2012), 1303–1310.

[199] Matthew J Vowels. 2021. Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods* (2021).

[200] Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review* 14, 5 (2007), 779–804.

[201] Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review* 25, 1 (2018), 35–57.

[202] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[203] Larry Wasserman. 2004. Bayesian inference. In *All of Statistics*. Springer, 175–192.

[204] Gary L Wells and Paul D Windschitl. 1999. Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin* 25, 9 (1999), 1115–1125.

[205] Shimon Whiteson, Brian Tanner, Matthew E Taylor, and Peter Stone. 2011. Protecting against evaluation overfitting in empirical reinforcement learning. In *2011 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*. IEEE, 120–127.

[206] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23, 1 (1996), 69–101.

[207] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903* (2021).

[208] Chhavi Yadav and Léon Bottou. 2019. Cold case: The lost mnist digits. *Advances in Neural Information Processing Systems* 32 (2019).

[209] Tal Yarkoni. 2022. The generalizability crisis. *Behavioral and Brain Sciences* 45 (2022).

[210] Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 12, 6 (2017), 1100–1122.

[211] Ed Yong. 2012. A failed replication draws a scathing personal attack from a psychology professor. *Discover* (2012). https://web.archive.org/web/20120313012842/http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/

[212] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 3 (2021), 107–115.

[213] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).