The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning

Jessica Hullman
Northwestern University
jhullman@northwestern.edu

Sayash Kapoor
Princeton University
sayashk@princeton.edu

Priyanka Nanayakkara
Northwestern University
priyankan@u.northwestern.edu

Andrew Gelman
Columbia University
gelman@stat.columbia.edu

Arvind Narayanan
Princeton University
arvindn@cs.princeton.edu

ABSTRACT

Arguments that machine learning (ML) is facing a reproducibility and replication crisis suggest that some published claims in research cannot be taken at face value. Concerns inspire analogies to the replication crisis affecting the social and medical sciences. A deeper understanding of what reproducibility concerns in supervised ML research have in common with the replication crisis in experimental science puts the new concerns in perspective, and helps researchers avoid "the worst of both worlds," where 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 causal attribution as exemplified in psychology versus predictive modeling as exemplified in ML. We identify common themes 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 difficult to refute due to underspecification of key parts of the learning pipeline. We conclude by discussing risks that arise when sources of errors are misdiagnosed and the need to acknowledge the role of human inductive biases in learning and reform.

CCS CONCEPTS

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

KEYWORDS

Machine learning, replication, science reform, generalizability.

ACM Reference Format:
Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. 2022. The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3514094.3534196

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534196

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—*generalizability or robustness*—where methods may work well under certain conditions but fail when applied to new problems or in the world [169], 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 [202] has inspired subsequent research proposing a variety of explanations for the apparent brittleness of performance (e.g., [54, 79, 96]). Terms like "alchemy" [123] 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 [3, 99, 138] and can involve data filtering decisions that impact achievable accuracy [33, 137, 138].

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, corrective actions should be taken, and claims discounted—especially when they cannot be externally reproduced [42, 85, 125].

One way that authors call attention to concerns in ML research is analogizing them to the replication crisis in psychology [19, 36, 110, 115]. While psychologists discussed fundamental issues with conventional approaches to inference as early as the 1960s [50, 148], in the last decade critics brought concerns to the forefront, demonstrating how motivated researchers can obtain false positives under various conditions [75, 76, 192] and that many published conclusions about human behavior in psychology research cannot be replicated [67, 161]. These revelations spur hard questions about what are necessary conditions for science, how to resolve uncertainty about published claims, and how to shift incentives.

Despite their focus on predictive modeling (and abundant recent successes in terms of performance on benchmarks, e.g., [71] and adoption in real-world applications, e.g., [145]), fields like AI and ML could learn from psychology's ongoing attempts to diagnose sources of non-replicability and reform conventional use and reporting of methods in the causally-focused explanatory modeling paradigm prevalent in psychology.[1] Taking a wider perspective on learning failures is well aligned with the idea of integrative modeling, referring to approaches that combine aspects of both paradigms [114, 115, 177, 224]. For example, social scientists might use prediction along with explanation to reduce overfitting to noisy experiment results, while researchers in fields like ML can incorporate explanatory methods to ascertain what a model appears to have learned. Integrative modeling acknowledges how researchers frequently misunderstand the relationship between explanation and prediction, assuming, e.g., that models that succeed in explaining have greater predictive validity [224] than those that appear less psychologically plausible ([106, 191] as cited in [224]) or that a model that achieves high predictive accuracy won't deviate much from what a human considers to be a plausible decision rule [79]. But to avoid integrative modeling leading to "the worst of both worlds," researchers will need to understand subtle differences in ways in which inferences can be limited in each paradigm. To date, connections that authors have drawn between these two reform discussions have been piecemeal, leaving it unclear what lessons, if any, might be gained from putting these domains in conversation.

To address this gap, we contribute a detailed comparison of limitations of inference in causally-driven explanatory versus predictive modeling. Our analysis is synthesized from 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. While we surface issues that affect various areas in psychology and ML, we ground our discussion around examples from experimental social psychology, which like ML relies on data reflecting human behavior and uses controlled comparisons to produce claims, and empirical research in supervised discriminative learning (i.e., classification) methods, including deep neural nets (DNNs) that encapsulate many recent concerns.

Our results highlight where concerns across the two domains can stem from similar types of oversights, including overreliance on 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 reason dichotomously about empirical results. 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 undisclosed sources of variance in the learning pipeline. We argue in particular that many of the errors recently discussed in ML expose the cracks in long-held beliefs that optimizing predictive accuracy using huge datasets absolves one from having to consider a true data generating process or formally represent uncertainty in performance claims. At the same time, the goals of ML are inherently oriented toward addressing learning failures, suggesting that lessons about irreproducibility could be resolved through further methodological innovation in a way that seems unlikely in social psychology. This assumes, however, that ML researchers take concerns seriously and avoid overconfidence in attempts to reform. We conclude by discussing risks that arise when sources of errors are misdiagnosed and the need to acknowledge the role that human inductive biases play in learning and reform.

[1] Also known as attribution [73], and typically involving estimation of regression surfaces and assignment of significance to individual predictors.

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 policy goals such as estimating effectiveness among different types 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 using interventions thought to interact with the target phenomena. 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 **model representation**, i.e., the model class or functional form that 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 more DNN architecture in ML (with a specific configuration of network parameters like arrangement into convolutions, activation functions, etc.). Because quantifying and searching *all* DGPs implied by probability distributions over the observation space tends to be prohibitively complex, learning pipelines often consider a subset or "small world" of model configurations [27], called the hypothesis space of the learner in ML. **Model selection** or model-based inference describes how a best fit model that is most consistent with the data is determined. 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 maximum likelihood estimation (MLE), etc.).

An **evaluation** may follow to validate the usefulness of what is learned 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. By "errors in learning," we refer to issues that arise in this larger process in which a researcher specifies and "solves" a learning problem.

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 [148, 191, 224]. Researchers identify hypotheses representing predictions about variables that constitute observed data. Often these constitute "weak theories" [149], 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's p-value 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" [84, 113] implied by authors to be generally true about human behavior. For example, thinking about old age induces old-like behavior [13].

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 (SOTA) in subareas (e.g., natural language processing (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 evaluation metric to be used as a scoring function (e.g., accuracy) [138]. 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 in empirical research papers typically report performance of a new learner (i.e., model) on benchmarks, compared to baselines representing the prior SOTA. 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 choosing a representation and a model selection and evaluation approach), and communication of results in a research paper.

3.1 Data collection and preparation

Social psychology. *High measurement error relative to signal, unacknowledged flexibility in defining data inputs, underspecified or non-representative subject samples, and underspecification of stimuli generation, and other "design freedoms" can threaten the validity of conclusions drawn in empirical psychology research.*

The design of many psychology experiments implies that researchers do not grasp the implications of using small samples and noisy measurements to draw inferences about effects that are a priori likely to be small. For example, a thought-to-be 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 [42, 192]. In reality, with a lower powered study, not only is there a lower probability of finding a true effect of a given size, but there is a lower probability that an observed effect which passes a significance threshold actually reflects a true effect that will appear under replication [42]. Under low power, estimates of observed effects will tend to reflect sampling error that derives from the limited size of the sample relative to a target population, and forms of measurement error [141], 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 to any plausible effect size [91].

Inherent flexibility in how a researcher specifies an analysis is a different type of threat. A "researcher degrees of freedom" or "garden of forking paths" metaphor [88, 192] suggests that human tendencies toward self-serving interpretations of ambiguous evidence (e.g., [11, 58] as cited in [192]), make researchers likely to draw conclusions that verify their hypotheses. Given an outcome of interest (e.g., self-reported political preference), an analyst may bias results toward a preferred conclusion by selecting data transformations and outlier removal processes, or choosing between different predictor variables or ways of operationalizing the outcome variable, conditional on seeing the results of these choices, without necessarily recognizing they are doing anything improper. 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, resulting in a form of procedural overfitting.

Scholars have pointed to study results not being reproducible because they use non-representative samples of a target population, such as convenience samples of university students from western educated industrialized rich democratic (WEIRD) countries [112]. As researchers have become more accustomed to the importance of statistical power and representative samples, online recruitment of participants in social psychology [183] has increased. However, it is unclear that sample homogeneity is addressed by online samples [46] and this trend has led to greater use of self-reported measures [183] 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 [151]. Another fundamental but often overlooked issue concerns how psychologists often leave the target population of their inferences unspecified [93], making it ambiguous what is being learned at all.

Machine learning. *Standardization of benchmarks and the prohibitive cost of amassing large datasets means that researchers often rely on existing datasets [108, 200], typically obtained through crowdsourced annotation and web-scale data (e.g., [61, 135]). Similar to psychology, factors like choosing how to transform data after seeing results, the use of non-representative samples, and underspecification of the population captured in data threaten the validity of claims. More frequently discussed issues include the differential effects of non-random measurement error on real world outcomes when a model is deployed and the way that a "good" predictive model can perpetuate forms of historical bias like stereotypes.*

Recent work in ML points 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 [137] as cited in [138]).

Non-representative samples are also a concern, including violations of the assumption that the development distribution from which the training and test data are presumed to be randomly drawn is the same as the deployment distribution from which samples will be drawn in real-world applications [14, 165, 201]. 'Representation bias" [201] involves 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., [41, 164, 227]). Suresh and Guttag [201] 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 describe how error in the (often unreported [78]) labeling process used to construct ground truth can lead to overfitting [38, 160], as well as how data preparation steps lose information whenever majority-rule is used to construct a ground truth without preserving information about label distributions (e.g., describing variance across annotators) [57, 98].

However, criticism of data practices in ML often focuses on systematic measurement error (i.e., bias) in collected data that threatens construct validity: whether the measurement is actually capturing the intended concept. "Measurement bias" [201] has been used to refer to differential measurement error [211], 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., [134] as cited in [201]). Jacobs and Wallach [126] 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.

A novel concern about measurement bias in ML relative to psychology occurs when biased input data are used to train a model and contribute to undesirable social norms. Data may record historical biases [201] (e.g., training a model to recognize successful applicants on data where women were admitted less due to bias). "Harms of representation" [1, 52] 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 [132]. The fact that ML is often intended for prescriptive use in the world, rather than descriptive use as in psychology research helps explain the prevalence of these concerns and the emphasis on systematic measurement error.

Finally, data concerns in ML increasingly refer to forms of underspecification of population details and underacknowledgment of the constructed nature of data, instead taking data as given [20, 78, 122, 186]. 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 [165].

3.2 Model representation

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

Social psychology. *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 incorporate prior knowledge into modeling informally rather than explicitly.*

When modeling a latent psychological phenomenon, often via simple measures of correlation and linear parametric models [30, 31], researchers implicitly assume that there is a true DGP that exactly captures how the target 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 [89]. 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 [213].

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 [26]. 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 [213]. Under these conditions, the validity of conventional interpretations of fitted models depends on the observational model faithfully approximating the true DGP [89].

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 [213]. 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, 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 [7] or including variables that could in

	Social Psychology	Machine Learning
Data selection and preparation	◦ High measurement error relative to the size of effects being studied [15, 62, 141] ◦ Data transformations decided contingent on (NHST) results [88, 192] ◦ Non-representative [112, 151] or underdefined samples [93]; insufficient stimuli sampling [92, 218, 223] ◦ Small samples and noisy measurements (low power) leading to biased estimates [42]	◦ Differential measurement error (e.g., across social groups) [41, 164, 201, 227] which is not modeled [126, 134] ◦ Label errors [38, 160] and disagreement [57, 98] ◦ Data transformations decided contingent on performance comparisons [33, 137] ◦ Underrepresentation of portions of input space in training data [14, 165, 201] ◦ Input data too huge to understand [20, 165]
Model representation	◦ Overreliance on models and designs with good asymptotic guarantees [159] ◦ No explicit representation of prior/domain knowledge [80, 89] ◦ Inappropriate expectations [49, 82, 223] in light of crud factor [149, 162]; belief in many nudging factors with large consistent effects on outcome [209] ◦ Unacknowledged multiplicity of solutions [224] ◦ Structural misspecification [144, 213]	◦ Overreliance on asymptotic (worst-case) guarantees [65] ◦ Underspecification of desired inductive biases [54, 124]; failure to prevent shortcut learning [79] ◦ Inappropriate i.i.d. assumption in light of real-world nonstationarity [29, 198, 220] ◦ Reliance on fine-tuning/foundation models for which hyperparameter tuning is opaque [64, 221] ◦ Convergence in architectures around large models [20, 32, 199]
Model selection and evaluation	◦ Implicit optimization for statistical significance [85, 87, 88, 97, 121, 136] ◦ Inference as black box [93, 136, 217]; Not motivating choice of estimator or optimization for particular inference goal [24, 215] ◦ Misunderstanding/misusing ideas of statistical significance [81, 102, 120, 121, 214] ◦ Multiple comparisons problem [86]	◦ Implicit optimization to beat SOTA [114, 188] ◦ Knowledge of how OOD test sets are constructed used to choose representation/method [207] ◦ Overlooked sensitivity of optimizer performance to hyperparameters [36, 47, 94, 168, 187]; computational budget [206] ◦ Presence of implementation variation [138] and tricks [6, 111] ◦ Misuse of cross validation [25, 45, 109, 116, 207] ◦ Optimism of cross validation [72, 146] ◦ Loss metric misalignment [119] ◦ Not comparing to simpler baselines [53, 188] or priors [101]
Communication of claims	◦ Unwarranted speculation about what evidence a p value provides [204] ◦ Overgenerationalization (i.e., beyond studied population) [60, 105, 178, 194, 223] ◦ Unavailable data and code [83, 117, 153] ◦ Not acknowledging having explored multiple analyses conditioned on data [86, 192] ◦ Inaccurate descriptions of what p values mean [4, 28, 90, 204]	◦ Unwarranted speculation about causes [131, 138, 140] ◦ Implying equivalence of learning problems and human performance on a task [131, 138, 140] ◦ Lack of dataset documentation [20, 78, 165] ◦ Inaccessible data, code, computational resources [99, 182, 197] ◦ Not reporting implementation conditions/sources of variance [140, 188] ◦ Underpowered performance comparisons [3, 36]; ignoring sampling error [3, 138, 173]

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

principle be affected by experimental manipulations (and hence represent outcome variables themselves [51]) cause the conventional interpretation of the fitted parameter values not to hold. However, researchers seldom acknowledge these limitations.

Researchers often choose designs based on a preference for simpler models. Perhaps the most common example is preferring between-subject designs 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 carryover effects between treatments experienced by the same individual [159]. However, high variation between people can lead to poor estimates of average treatment effects if the treatment interacts with background variables associated with differences in individuals and contexts [144, 159].

More generally, psychology researchers have been criticized for estimating effects as if they are constant rather than assuming they will vary across people or contexts [82]. 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 [223] (e.g., a "fixed effect fallacy" [49]).

Tendencies to overlook important sources of variation in modeling are implied by Meehl's conception of the "crud factor" [149, 162], which emphasizes how causal attribution using constrained model spaces 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 [162]. Problems arise when researchers overlook model misspecification due to conventional but questionable beliefs about reality. For example, a tendency toward reporting model fits suggesting that novel yet seemingly trivial "nudging" factors (e.g., whether or not someone is menstruating [70] or whether there was a recent shark attack [2]) 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 [209].

In this way, choices of model representation (i.e., low dimensional linear regressions) are not fully consistent with prior knowledge. Conventional approaches to estimating an effect of interest are also memoryless in the sense that even when prior estimates of an effect of interest are available, e.g., from past experiments, they are generally not incorporated in the model representation. Combined with incentives to publish surprising results [74, 185] 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. *In theory, optimizing for predictive accuracy does not require well-approximating a true DGP. However, researchers' commonly assume that unseen data are drawn from the same distribution as training data and use asymptotics to motivate model choice, leading to unrealistic beliefs about the predictability of real world processes. Threats also arise from failures to explicitly represent a priori human expectations about what predictors are valid for a task, and a convergence on hard-to-analyze models that combine pre-trained representations with domain-specific data.*

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" [181] in the sense that the true DGP is in the set of learnable functions (or hypothesis space), nor even that the fitted function approximates the structure of the true DGP. Instead, the goal of learning can be formulated as identifying a function with error that can be guaranteed to fall within some bound of the best possible predictor over possible samples [210]. Choosing a representation (i.e., hypothesis space) in theory means reasoning about the inductive biases (i.e., properties of the predictors) it will return in light of prior knowledge, but in practice theoretical guarantees (e.g., worst-case bounds) on convergence or generalization ability have driven representation choices. This can lead, as in psychology, to use of models in cases where asymptotic assumptions don't apply [65]. Or, as in the case of much recent DL research, a more empirical, performance-driven approach uses achievable performance as the primary driver of model choice [174].

One of the most commonly cited deficiencies attributed to model representations in applied ML involves assuming a static relationship between the predictor variables and the outcome [212], which supports conventions like shuffling input data to create training and test sets [9]. This assumption makes models vulnerable to concept drift [220] (a.k.a. covariate shift [29] or distribution or dataset shift [198]), 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., [143]), behavioral reactions (e.g., [167]), or other unforeseen dynamics [170]. Under conventional "distribution unawareness," it also becomes difficult to distinguish when unexpected errors arise from distribution shift versus inefficiencies in the learning pipeline [23]. Distribution shift can lead to poorly calibrated estimates of the uncertainty of model performance [163], 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 greater focus on how different models fare at out-of-distribution (OOD) error and their robustness to adversarial manipulation, i.e., small changes to an input in feature space that dramatically change the predicted output (e.g., [16, 44, 179, 202, 207]). Recent results related to adversarial nonrobustness [124], underspecification [54], shortcut learning [79], simplicity bias [190], and competency problems [77] suggest that beliefs about the true DGP in predictive modeling as in ML are not necessarily as distinct from explanatory, attribution-oriented modeling as past comparative accounts (e.g., [39]) 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. [79] use "shortcut learning" to refer to a tendency for ML models to learn simple decision rules (e.g., [10, 129, 147]) that perform well on standard benchmarks. While these features represent "real" correlations, 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 might naturally expect successful performance to require 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 [79], 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 [54, 79, 124].

A related theory representing a symptom of predictive multiplicity is underspecification [54]: 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 [54]. Underspecification is distinct from forms of distribution shift that may give rise to shortcut learning, such as the presence of 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 [180].

A common approach to overcoming poor generalization of a model is to combine multiple 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 [22]. "Fine-tuning" pretrained "foundation" models [32] for domain-specific applications has become standard practice based on the performance that can be achieved over conventional domain-specific learning pipelines [104, 200, 221]. Though training on highly diverse input data tends to provide foundation models with inductive biases that improve extrapolation, a challenge is that fine-tuning performance can be highly sensitive to how poorly-understood parameters are set, making results hard to replicate [64]. For example, the robustness of a fine-tuned model has been found to vary considerably under small changes to hyperparameters [221]. Related is a concern

that the convergence in deep learning research around large DNN model architectures with minimal task-specific parameters [32] doubles down on an approach that imposes unreasonable environmental [20, 199] and research opportunity costs [20].

More generally, understanding the implications of model selection is complex for DNNs, where classical theory falls short of explaining the generalization performance (e.g., [17, 18, 55, 127, 226]). This has motivated lines of theoretical work that explore different explanations of phenomena like "double descent" [17], where the generalization performance of a deep model continues to improve even after it has achieved zero loss on (or perfectly interpolated) the training data. For example, some analyze the properties of over-parameterized linear regressions [55].

3.3 Model selection and evaluation

Model-based inference involves explicit and implicit choices of objective function, optimization approach, and evaluation metric.

Social psychology. *Claims made in social psychology research are threatened when researchers treat conventional approaches to model-based inference as a black box for consuming data and outputting inferences [12, 93, 136], and by researchers' implicit use of statistical significance alone as a criterion for deciding what to report.*

It is relatively rare for psychology research contributions to include explicit motivation for the estimators and loss functions used in modeling. Such "inference by convention" can produce misleading claims without outright cheating or motivated reasoning similar to how blindly preferring between-subjects designs can. For example, conventional use of maximum likelihood estimators based on their consistency [217] may lead researchers to overlook critical assumptions required for these estimators to be well-calibrated (i.e., have sampling distributions which are asymptotically normal). Analytical approaches to optimization bring convenience, but commonly-used approaches to model fitting and selection tend to be based on pre-experimental guarantees (i.e., before data are collected), which cannot guarantee that they will be appropriate (e.g., well-calibrated) on a particular dataset [24, 215].

A different source of misleading claims is the use of statistical significance as a coarse objective function. Implicit optimization for significance, in which researchers are essentially searching through a garden of forking paths for specifications that achieve significance as a sort of quasi-optimization approach [88], means that conventional interpretations of fitted models and statistical tests on parameter estimates will not hold. For example, the multiple comparisons problem, in which researchers neglect to control for data-dependent selection in what they report, alters the statistical properties of estimates and tests [86]. At the highest level, bias affects the published record when researchers decide whether to report results based on the significance levels [85, 185].

Using "statistical significance" as an implicit objective does not line up with scientific goals (e.g., [81, 102, 120, 121, 214]). For example, 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 [97, 121, 136]. In other words, hypothesis testing can sometimes be used as a sort of "truth mill" in psychology [85, 87].

Related problems include not acknowledging that as a random variable, p can vary considerably even under idealized replication [34, 90, 97, 154, 189], 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 [5, 103, 171].

Machine learning. *Claims are often made in ML research without acknowledging that they depend critically on choices of hyperparameters, initial conditions, and other configuration details that directly influence performance in non-convex optimization. Researchers also may exploit flexibility in designing performance comparisons in order to achieve superior performance for their contributed approach relative to alternatives [114, 188].*

In contrast to loss functions in simple regression models, ML models tend to have high dimensional non-convex loss. While this does not necessarily prevent generalization [48] it makes solutions like saddle points, which can give the illusion of a satisfactory local minimum, of greater concern [56, 184]. 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, to make non-convex optimization tractable requires setting various opaque hyperparameters and initial conditions that influence how the loss landscape is traversed [47, 94, 187].

For an optimization approach like stochastic gradient descent (SGD), hyperparameters like 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 and it may require too long [36]. Adaptive optimizers (e.g., Adagrad, Adam) allow hyperparameters like learning rate to vary for each training parameter, inducing a new dynamical system with each run and complicating attempts to explain what parts of a pipeline improved performance.

Hyperparameter tuning is also a computationally expensive task [206], inducing uncertainty about how a solution might differ under a larger computational budget or different parameter settings. 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 [187]. Similarly, sufficient hyperparameter optimization can mostly eliminate claimed performance differences in generative adversarial networks (GANs) [142], and better hyperparameter tuning on baseline implementations can eliminate evidence of performance advantages of new learning methods [111, 150]. Liao et al. [138] 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" [6, 111]—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 [133, 155, 219].

Other inference concerns pertain to 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 aims to identify proxies

for properties like learnability and generalizability (e.g., [127, 226]). Recent results show how counter to classical expectations about overfitting, minimizing training error without explicit regularization over overparameterized models tends to result in good generalization [158, 195, 226], driving a new theoretical agenda aimed at disentangling optimization methods and statistical properties of the solutions they find. Some emergent properties have been criticized. Related to shortcut learning, SGD has been shown to exhibit "simplicity bias"—a preference for learning simple predictors first, resulting in neural nets relying exclusively on the simplest features, for example, image color and texture, and remaining invariant to complex predictive features, for example, object shape [130, 190].

Other concerns with external validity arise when an explicitly chosen objective function is not a good proxy for the metric of interest in using the models, called "loss-metric misalignment" and threatening generalization [119]. For example, cross-entropy loss is often used as a loss function, whereas the evaluation metric of interest is often classification error or AUCPR. More generally, reporting single scalar error measures by convention overlooks important error variation (e.g., [69]).

Internal and external validity is threatened by leakage—broadly, using information from the test data in training—paralleling the reuse of data for choosing and evaluating a model's fit in psychology. Leakage can arise when a single CV procedure is used for model tuning and estimating error at once [45, 109, 116]. Failure to carefully consider which steps involved in training should be performed on each fold during CV can bias error estimates on test data [109], as can contaminating the procedure with future data in time series applications [25]. More generally, using 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 [72, 146], not unlike how low-power experiments lead to overestimates of effects in psychology.

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 [43] 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 [99].

Finally, implicit optimization for good performance results can also occur in ML. Improving performance on benchmark datasets, which have been thought to have caused most major ML research breakthroughs in the last 50 years [66], is how researchers showcase improvement in model performance to get published in top conferences and journals [172, 188] across ML subareas (e.g., [61, 118, 216]). 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 (Section 4), 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 [101], performs in a popular visual question answering task (VQA) [8]). Attempts to use OOD data to improve task performance are not valid when researchers rely on explicit knowledge of how the OOD splits were constructed or use the OOD test set for model validation [207].

4 COMMUNICATION OF CLAIMS

Sources of error can remain unacknowledged due to communication norms that suppress uncertainty and limit reproducibility.

Social psychology. *The contribution of a social psychology experiment can be framed as a stylized fact: a statement presumed generally true and replicable [84, 113] about some aspect of the world. Results are used to motivate broad claims [60], with deficiencies attributable to authors failing to acknowledge exploration of multiple analysis paths contingent on the data, and tending to downplay inherent dependencies and uncertainty when describing results.*

Because stylized facts derive from the results of experiments in laboratory-like environments, often on non-representative samples [84], credible reporting would emphasize the specific conditions studied [194]. Instead, however, researchers routinely state their findings in broad terms in articles, referring to how an intervention or trait affects "people" or entire groups [60, 105, 178], not acknowledging potential variation untested by theory and data.

Authors can perpetuate p-value fallacies when they write about effects as if present or absent (e.g., [28]) or overinterpret alternative hypotheses [204]. Or they may imply that a lack of significance is evidence of an absence of effect [4, 28] or that there is a significant difference between significant and non-significant results [90].

Finally, while sharing of data and analysis code has increased in psychology in recent years, many authors have not adopted such sharing (e.g., [205]). 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 [83, 117, 153].

Machine learning. *Communication concerns in ML include tendencies to not report 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.*

ML researchers often report point estimates of performance without quantifying uncertainty [3, 138, 173] or reporting key inputs 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 [140]. As examples, authors often do not report the number of models trained and the negative results found before the one they highlight is selected [3, 188]. Authors may cut corners since computing uncertainty and variance in ML models can incur significant computational costs, especially for large ML models [3, 36]. 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 [99], the computational budget [63], the choice of hyperparameter values, and the random initialization of ML models [47, 142, 187], 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 [157].

As with psychology, researchers may be tempted to speculate about causes without couching them in speculative terms [140]. 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 a more specific definition of the problem [138, 140] for which publishable results were achieved. For example, using "reading comprehension" to refer to a process is misleading when the model may not have used what a human would call critical information, like the text it is "comprehending" [131] (Section 3.3). More broadly, claims about performance are rarely evaluated in the context of relevant real-world applications [138].

ML faces analogous issues to the lack of open data and code in social psychology [68, 99, 107, 182, 197]. Details about dataset limitations that can threaten external validity [20, 78, 165] are often unreported in ML literature (Section 3.1), perhaps because new techniques for model creation have historically been valued over documenting datasets [186]. As in psychology, checking computational reproducibility of results requires making the complete code and data available with published papers [40]. Recent work attempts reproducibility checklists, documentation checklists, community challenges, and workshops [78, 152, 169]. However, while assessing replication in the social sciences is not trivial (e.g., [196]), a somewhat unique challenge in ML is that with the creation and widespread use of large ML models requiring significant computational resources [20], especially in NLP tasks, it becomes impossible for many researchers to even attempt replicating certain results.

5 IMPLICATIONS: WHAT CAN WE LEARN FROM THIS COMPARISON?

As researchers move toward integrative modeling, they should grasp common blind spots; the summary in Table 1 roughly orders issues to emphasize where concerns overlap between the two fields.

We see evidence of different 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 or modeling other forms of variation, like in 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" [148] and improbability that multiple predictors thought to have large effects on the same outcome would not also correlate with one another [209]. 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 [55], but have been slow to part with i.i.d. assumptions.

Norms around what is publishable in each field incentivize researchers to hack results to meet implicit objectives such as statistical significance 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, so that results do not generalize as assumed. 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.

5.1 Irrefutable claims

On a deeper level, claims researchers are making in both fields 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 incentives to publish positive evidence on often counterintuitive effects [74, 185]), "false positive psychology" [192] is not a surprising result. Consider how much more difficult, 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, defined target populations, and attempts to model other sources of contextual variation, assuming that claims made about any particular parameter estimates obtained through analyzing experiment results generalize beyond that particular set of participants, stimuli, etc. is not credible.

Turning to ML, many reproducibility failures seem to derive from a similar tolerance for irrefutable contributions, manifesting 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 some accuracy on a benchmark dataset. Even if we assume the reported accuracy is not 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 dataset, because the population from which benchmark datasets are drawn are rarely specified to the detail needed for another sample to be drawn [128]. 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 [175]. 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 fragile [59, 208].

At a higher level, analogies between human and artificial intelligence are embedded in AI culture, but without specifying the neurocomputational processing involved in cognition [176], whether ML approaches capture key aspects of human consciousness (e.g., [100]) or new algorithms intended to instantiate human-like mechanisms (e.g., [21]) succeed in a human-like way is 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 [35, 59, 139], a lack of rigor in evaluation [138, 188], and overreliance on theory [95, 140] 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. This could naturally encourage a more cautious and empirical mindset among

researchers, but only if pressures to make bold claims from structural incentives that encourage "planting one's flag" before others do [188] don't outweigh the trend toward embracing uncertainty.

5.2 Latent expectations versus reality

Characterizing the conventions that give rise to irrefutable claims as forms of underspecification—meaning that some aspect of the learning problem has not been formalized to an extent that allows it to be solved—might help point researchers toward new methods to address what is missing. In particular, the role of human expectations in defining "success" in learning has been implicit, but innovations are often driven by making these expectations explicit.

Colloquially, many ML methods are assumed to free researchers from theorizing how well a fitted function captures critical structure in the true DGP. Yet, many weaknesses being identified suggest that the reality of non i.i.d. test data is pushing ML researchers in "purely" predictive areas toward philosophies underlying explanation. Recent definitions of underspecification [54], shortcut learning [79], adversarial vulnerability [124], etc. motivate the need to impose more constraints on what is learned, and the most natural source is the human who assesses and interprets the results.

In social psychology, DGPs are modeled, if only as a symptom of using conventional inference. However, we see a displacement of prior knowledge in designing and interpreting experiments, where a priori expectations about how big an effect could be are often overlooked. There is also a failure to acknowledge how the styles of research the field rewards, 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.

There is little reason to believe that taking steps toward integrative modeling will greatly improve practice if researchers fail to actively monitor what new methods help them learn for the issues listed in Table 1. The worst of both worlds would result if instead they assume that any use of 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."

Another bright spot in times of methodological crisis is that when mismatch is recognized, it can lead to new technical innovations. Paradoxically, striving to identify a single foolproof solution to a recognized learning problem can drive new techniques to close the gap between expectations and reality. For example, in ML recognizing the "brittleness" of deep learning models in light of human perceptions of learning has led to major improvements to generalization from new approaches to adversarial robustness. ML may have an advantage over psychology in addressing reproducibility problems in that a 'cat and mouse' dynamic characterizes many recent successes, where evaluative work is followed by clever changes to the definition of a learning problem or pipeline that overcome that weakness. Perhaps the most important question for modern AI and ML is what, if any, forms of mismatch between human expectations and model behavior cannot be solved through a reframing of the learning problem to find beneficial new "hacks."

5.3 Epistemological gaps and rhetorical risks

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.

Consider how irrefutable theories and claims induce greater dependence 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 always agree on what constitutes successful replication [166], and intuitions can be proven wrong. For example, under a formal definition of a study's reproduciblity rate, reproducing experimental results does not necessarily indicate a "true" effect, and vice versa for a "false" effect [62]. In ML, tests are similarly indirect, but the stakes often higher: when an approach fails to perform as expected in the world, researchers may scrutinize the original claims, but at the expense of those affected in deployment. Progress can be hard to judge due to the speed of discovery [37] and conflicting valuations of standard approaches like benchmarks (e.g., [172, 222]).

There is a need to accurately diagnose the fundamental problems, rather than 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 problems preregistration addresses point to the challenge of determining when a given reform should have privileged status [156, 193, 203].

Sociological issues also arise when researchers believe good generalization can be achieved via a singular universal method of statistical inference [93]. 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 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 [67]) and opposed by those whose work has failed to replicate (for example, [225]). This would lead to open-science practices themselves being a marker of research quality. On the other hand, honesty and transparency are not enough [83]: openness and preregistration alone won't endow replicability to a psychology study with a high ratio of noise to signal, which can happen when designs focus on procedural issues (e.g., randomization), to the detriment of theory and measurement.

Steps can be taken to reduce rhetorical risks. For example, Devezer et al. [62] propose accompanying colloquial statements about reproducibility problems and solutions with formal problem statements and results, and provide questions to guide researchers in doing so. Greater rigor in reform arguments can mean quicker identification of logical errors, misintepretations of constructs, or other blind spots in attempts to steer a field back on track.

6 ACKNOWLEDGMENTS

We thank Jake Hofman, Cyril Zhang, Jean Czerlinski Oretga for comments. Hullman is supported by the NSF (IIS-1930642) and a Microsoft Faculty Fellowship, Narayanan by the NSF (IIS-1763642), and Gelman by the ONR.

REFERENCES

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

[2] Christopher H Achen and Larry M Bartels. 2012. Blind retrospection: Why shark attacks are bad for democracy. *Center for the Study of Democratic Institutions, Vanderbilt University. Working Paper* (2012).

[3] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *NeurIPS* 34 (2021).

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

[5] 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.

[6] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. 2020. What matters for on-policy deep actor-critic methods? A large-scale study. In *ICLR*.

[7] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton university press.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proc. of ICCV*. 2425–2433.

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

[10] 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.

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

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

[13] John A Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *J. of Personality and Social Psychology* 71, 2 (1996), 230.

[14] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[15] 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.

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

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

[18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. 2018. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *NeurIPS* 31 (2018).

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

[20] 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 *Proc. of FAccT*. 610–623.

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

[22] 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.

[23] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. 2020. Cats are not fish: Deep learning testing calls for out-of-distribution awareness. In *Proc. of ASE*. 1041–1052.

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

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

[26] Jose M. Bernardo and Adrian F. M. Smith. 1994. *Bayesian Theory*. Wiley.

[27] 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/

[28] 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.

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

[30] 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.

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

[32] 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:2108.07258* (2021).

[33] Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. 2015. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. of autism and developmental disorders* 45, 5 (2015), 1121–1136.

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

[35] 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)*.

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

[37] Samuel Bowman. 2022. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. In *Proc. of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*. 7484–7499.

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

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

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

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] 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:1910.05446* (2019).

[48] 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.

[49] Herbert H Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. of Verbal Learning and Verbal Behavior* 12, 4 (1973), 335–359.

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

[51] 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, Alan E Hubbard, and Mark J van der Laan. 2020. Targeting learning: Robust statistics for reproducible research. *arXiv:2006.07333* (2020).

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

[53] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

[54] 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:2011.03395* (2020).

[55] 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:2109.02355* (2021).
[56] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *NeurIPS* 27 (2014).
[57] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the ACL* 10 (2022), 92–110.
[58] 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.
[59] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *arXiv:2107.07002* (2021).
[60] Jasmine M DeJesus, Maureen A Callanan, Graciela Solis, and Susan A Gelman. 2019. Generic language in scientific communication. *Proc. of the National Academy of Sciences* 116, 37 (2019), 18370–18377.
[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 248–255.
[62] 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.
[63] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. *arXiv:1909.03004* (2019).
[64] 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:2002.06305* (2020).
[65] Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 10 (2012), 78–87.
[66] David Donoho. 2017. 50 years of data science. *J. of Computational and Graphical Statistics* 26 (2017), 745–766.
[67] 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. *Proc. of the National Academy of Sciences* 112 (2015), 15343–15347.
[68] Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the ACL* 5 (2017), 471–486.
[69] Chris Drummond. 2006. Machine learning as an experimental science (revisited). In *AAAI Workshop on Evaluation Methods for Machine Learning*. 1–5.
[70] Kristina M Durante, Ashley Rae, and Vladas Griskevicius. 2013. The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science* 24, 6 (2013), 1007–1016.
[71] Peter Eckersley, Yomna Nasser, et al. 2017. EFF AI progress measurement project. *Retreived from: https://eff. org/ai/metrics, accessed on* (2017), 09–09.
[72] Bradley Efron. 2004. The estimation of prediction error: Covariance penalties and cross-validation. *J. of the American Statistical Association* 99, 467 (2004), 619–632.
[73] Bradley Efron. 2020. Prediction, estimation, and attribution. *International Statistical Review* 88 (2020), S28–S59.
[74] Daniele Fanelli. 2010. "Positive" results increase down the hierarchy of the sciences. *PloS one* 5, 4 (2010), e10068.
[75] Gregory Francis. 2012. The psychology of replication and replication in psychology. *Perspectives on Psychological Science* 7 (2012), 585–594.
[76] Gregory Francis. 2012. Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review* 19 (2012), 975–991.
[77] 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:2104.08646* (2021).
[78] 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.
[79] 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.
[80] Andrew Gelman. 2012. Ethics and statistics: Ethics and the statistical use of prior information. *Chance* 25, 4 (2012), 52–54.
[81] Andrew Gelman. 2013. P values and statistical practice. *Epidemiology* 24, 1 (2013), 69–72.
[82] Andrew Gelman. 2015. The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *J. of Management* 41, 2 (2015), 632–643.
[83] Andrew Gelman. 2017. Honesty and transparency are not enough. *Chance* 39 (2017), 37–39. Issue 1.
[84] 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.
[85] 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.
[86] 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).
[87] Andrew Gelman and Eric Loken. 2014. Ethics and statistics: The AAA tranche of subprime science. *Chance* 27, 1 (2014), 51–56.
[88] 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.
[89] 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.
[90] 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.
[91] 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.
[92] Gerd Gigerenzer. 2022. We need to think more about how we conduct research. *Behavioral and Brain Sciences* 45 (2022).
[93] Gerd Gigerenzer and Julian N Marewski. 2015. Surrogate science: The idol of a universal method for scientific inference. *J. of Management* 41, 2 (2015), 421–440.
[94] Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Dahl, Zack Nado, and Orhan Firat. 2021. A loss curvature perspective on training instabilities of deep learning models. In *ICLR*.
[95] 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
[96] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv:1412.6572* (2014).
[97] Steven N Goodman. 1993. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American J. of Epidemiology* 137, 5 (1993), 485–496.
[98] 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 *Proc. of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[99] Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proc. of ACL*. 2786–2791.
[100] Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *arXiv:2011.15091* (2020).
[101] 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 *Proc. of CVPR*. 6904–6913.
[102] 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.
[103] Sander Greenland and Zad Rafi. 2019. To aid scientific inference, emphasize unconditional descriptions of statistics. *arXiv:1909.08583* (2019).
[104] 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:2004.10964* (2020).
[105] Kris D Gutiérrez and Barbara Rogoff. 2003. Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher* 32, 5 (2003), 19–25.
[106] Michael R Hagerty and V Srinivasan. 1991. Comparing the predictive powers of alternative multiple regression models. *Psychometrika* 56, 1 (1991), 77–85.
[107] 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.
[108] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems* 24, 2 (2009), 8–12.
[109] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.
[110] Will Douglas Heaven. 2020. AI is wrestling with a replication crisis. *MIT Technology Review* (2020).
[111] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proc. of the AAAI*, Vol. 32.
[112] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain sSiences* 33, 2-3 (2010), 61–83.

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

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

[115] 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.

[116] Mahan Hosseini, Michael Powell, John Collins, Chloe Callahan-Flintoft, William Jones, Howard Bowman, and Brad Wyble. 2020. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews* 119 (2020), 456–467.

[117] 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.

[118] 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* (Feb. 2021). http://arxiv.org/abs/2005.00687 arXiv: 2005.00687.

[119] 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.

[120] Raymond Hubbard and MJ Bayarri. 2003. P values are not error probabilities. *Institute of Stat. and Dec. Sci., Working Paper* 03-26 (2003), 27708–0251.

[121] 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.

[122] 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 *Proc. of FAccT*. 560–575.

[123] Matthew Hutson. 2018. Has artificial intelligence become alchemy? *Science* (2018).

[124] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. *NeurIPS* 32 (2019).

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

[126] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proc. of FAccT*. 375–385.

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

[128] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proc. of FAccT*. 306–316.

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

[130] 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. *NeurIPS* 32 (2019).

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

[132] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.

[133] Khimya Khetarpal, Zafarali Ahmed, Andre Cianflone, Riashat Islam, and Joelle Pineau. 2018. RE-EVALUATE: Reproducibility in evaluating reinforcement learning algorithms. *2nd Reproducibility in ML Workshop (ICML)* (2018).

[134] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 108. 22–27.

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

[136] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

[137] Thomas Liao, Benjamin Recht, and Ludwig Schmidt. 2020. In a forward direction: Analyzing distribution shifts in machine translation test sets over time. *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*.

[138] 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)*.

[139] 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

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

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

[142] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are GANs created equal? A large-scale study. *NeurIPS* 31 (2018).

[143] 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:2111.07408* (2021).

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

[145] Roger Magoulas and Steve Swoyer. 2020. *AI Adoption in the Enterprise*. Beijing: O´ Reilly. Recuperado de http://www. oreilly. com/data/free/ai ....

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

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

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

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

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

[151] Xiao-Li Meng. 2018. Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics* 12, 2 (2018), 685–726.

[152] 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 *Proc. of FAccT*. 220–229.

[153] 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.

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

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

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

[157] 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.

[158] 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:1412.6614* (2014).

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

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

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

[162] Amy Orben and Daniël Lakens. 2020. Crud (re) defined. *Advances in Methods and Practices in Psychological Science* 3, 2 (2020), 238–247.

[163] 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. *NeurIPS* 32 (2019).

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

[165] 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.

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

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

[168] David Picard. 2021. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv:2109.08203* (2021).

[169] 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. *J. of Machine Learning Research* 22 (2021).

[170] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

[171] 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.

[172] 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:2111.15366* (2021).

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

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

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

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

[177] Roberta Rocca and Tal Yarkoni. 2021. Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science* 4, 3 (2021), 25152459211026864.

[178] Barbara Rogoff. 2003. *The Cultural Nature of Human Development*. Oxford University Press.

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

[180] 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.

[181] Stuart J Russell and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*.

[182] 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.

[183] 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.

[184] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120* (2013).

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

[186] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. of CSCW* 5 (2021), 1–37.

[187] 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.

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

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

[190] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. *NeurIPS* 33 (2020), 9573–9585.

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

[192] 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.

[193] 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. *J. of Consumer Psychology* 31, 1 (2021), 177–180.

[194] 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.

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

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

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

[198] Amos Storkey. 2009. When training and test sets are different: Characterizing learning transfer. *Dataset Shift in Machine Learning* 30 (2009), 3–28.

[199] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proc. of the AAAI*, Vol. 34. 13693–13696.

[200] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proc. of ICCV*. 843–852.

[201] 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.

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

[203] Aba Szollosi, David Kellen, Danielle Navarro, Richard Shiffrin, Iris van Rooij, Trisha Van Zandt, and Chris Donkin. 2020. Is preregistration worthwhile? *Trends in Cognitive Sciences* 24 (2020), P94–95. Issue 2.

[204] Denes Szucs and John Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience* 11 (2017), 390.

[205] 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.

[206] 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.

[207] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. 2020. On the value of out-of-distribution testing: An example of Goodhart's law. *NeurIPS* 33 (2020), 407–417.

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

[209] 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:2105.13445* (2021).

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

[211] 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 J. of Epidemiology* 175, 12 (2012), 1303–1310.

[212] Vladimir Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

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

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

[215] 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.

[216] 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 *Proc. of the EMNLP Workshop BlackboxNLP*. ACL, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

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

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

[219] Shimon Whiteson, Brian Tanner, Matthew E Taylor, and Peter Stone. 2011. Protecting against evaluation overfitting in empirical reinforcement learning. In *ADPRL*. IEEE, 120–127.

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

[221] 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:2109.01903* (2021).

[222] Chhavi Yadav and Léon Bottou. 2019. Cold case: The lost mnist digits. *NeurIPS* 32 (2019).

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

[224] 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.

[225] 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/

[226] 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.

[227] 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:1707.09457* (2017).