Position: Why We Must Rethink Empirical Research in Machine Learning

Moritz Herrmann [1 2] **F. Julian D. Lange** [1 2] **Katharina Eggensperger** [3] **Giuseppe Casalicchio** [4 2] **Marcel Wever** [5 2]
Matthias Feurer [4 2] **David Rügamer** [4 2] **Eyke Hüllermeier** [5 2] **Anne-Laure Boulesteix** [1 2] **Bernd Bischl** [4 2]

Abstract

We warn against a common but incomplete understanding of empirical research in machine learning that leads to non-replicable results, makes findings unreliable, and threatens to undermine progress in the field. To overcome this alarming situation, we call for more awareness of the plurality of ways of gaining knowledge experimentally but also of some epistemic limitations. In particular, we argue most current empirical machine learning research is fashioned as confirmatory research while it should rather be considered exploratory.

1. The Non-Replicable ML Research Enigma

In his Caltech commencement address "Cargo Cult Science"*,[1] Richard Feynman (1974) described how researchers employ practices that conflict with scientific principles to adhere to a certain way of doing things. ***This position paper warns against similar tendencies in empirical research in machine learning (ML) and calls for a mindset change to address methodological and epistemic challenges of experimentation.***

There is ML research that does not replicate. From an empirical scientific perspective, non-replicable research is a fundamental problem. As Karl Popper (1959/2002, p. 66) phrased it: "non-reproducible single occurrences are of no significance to science."[2] Consequently, ML research that does not replicate has far-reaching epistemic* *and* practical consequences. From an epistemological* point of view, it means that research results are unreliable and, to some extent, it calls into question progress in the field. In practice, it may jeopardize applied empirical researchers' confidence in experimental results and discourage them from applying ML methods, even though these novel approaches might be beneficial. For example, ML is increasingly being used in the medical domain, and this is often promising in terms of patient benefit. However, there are also examples indicating that applied researchers (are starting to) have concerns about ML being used in this high-stakes area. Consider, for example, this quite drastic warning by Dhiman et al. (2022, p. 2): "Machine learning is often portrayed as offering many advantages [...]. However, these advantages have not yet materialised into patient benefit [...]. Given the increasing concern about the *methodological quality and risk of bias of prediction model studies* [emphasis added], caution is warranted and the lack of uptake of models in medical practice is not surprising." That is, if the ML community does not improve rigor in empirical methodological research, we think there may be a risk of a backlash against the use of ML in practice.

In general, there is a growing body of empirical evidence showing that conclusions drawn from experimental results in ML were overly optimistic at the time of publication and could not be replicated in subsequent studies. For example, Melis et al. (2018, p. 1) "arrive at the somewhat surprising conclusion that standard LSTM architectures, when properly regularized, outperform more recent models"; Henderson et al. (2018, p. 3213) found for deep reinforcement learning "that both intrinsic (e.g. random seeds, environment properties) and extrinsic sources (e.g. hyperparameters, codebases) of non-determinism can contribute to difficulties in reproducing baseline algorithms"; Christodoulou et al. (2019, p. 12) found in a systematic review "no performance benefit of machine learning over logistic regression for clinical prediction models"; Elor & Averbuch-Elor (2022, p. 1) found in their study on data balancing in classification "that balancing does not improve prediction performance for the strong" classifiers; see also Lucic et al. (2018), Riquelme et al. (2018), Raff (2019), Herrmann et al. (2020), Ferrari Dacrema et al. (2021), Marie et al. (2021), Buchka et al. (2021), Narang et al. (2021), van den Goorbergh et al.

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich, Munich, Germany [2]Munich Center for Machine Learning (MCML), Munich, Germany [3]University of Tübingen, Tübingen, Germany [4]Department of Statistics, LMU Munich, Munich, Germany [5]Institute of Informatics, LMU Munich, Munich, Germany. Correspondence to: Moritz Herrmann <moritz.herrmann@lmu.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]As our paper contains some jargon, we have included a glossary in the appendix; asterisks (*) in the text denote covered terms.

[2]Reproducible here does not refer to exact *computational reproducibility** but generally to arriving at the same scientific conclusions, termed replicability* in this paper.

(2022), Mateus et al. (2023), McElfresh et al. (2023), or the surveys by Liao et al. (2021) and Kapoor & Narayanan (2023) for similar findings. In concrete terms, there is published ML research that is, as Popper would say, *of no significance to science*, but we do not know how much!

We have been warned; don't we listen? We are by no means the first to raise these and related issues, and the very fact that we are not the first is a matter of even graver concern. We think that empirical research in ML finds itself in a situation where practicing questionable research practices, such as *state-of-the-art-hacking* (SOTA-hacking; Gencoglu et al., 2019; Hullman et al., 2022), has sometimes become more rewarding than following the long line of literature warning against it. Langley wrote an editorial "Machine Learning as an Experimental Science" as early as 1988, and Drummond and Hand pointed out problems with experimental method comparison in ML already in 2006. Apart from these specific examples, there is a range of literature over the last decades dealing with similar issues (e.g., Hooker, 1995; McGeoch, 2002; Johnson, 2002; Drummond, 2009; Drummond & Japkowicz, 2010; Mannarswamy & Roy, 2018; Sculley et al., 2018; Lipton & Steinhardt, 2018; Bouthillier et al., 2019; Liao et al., 2021; D'Amour et al., 2022; Raff & Farris, 2022; Lones, 2023; Trosten, 2023). Specifically relevant is the paper by Nakkiran & Belkin (2022, p. 2), in which they note a "perceived lack of legitimacy and real lack of community for good experimental science" (still) exists. If we continue not to take these warnings seriously the amount of non-replicable research will only continue to increase, as the cited *very recent* empirical findings indicate. We do not believe that deliberate actions on the part of individuals have led to this situation but that there is a general unawareness of the fact that, while "follow[ing] all the apparent precepts and forms of scientific investigation [in ML]," one can be "missing something essential." In particular, this includes that "if you're doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked" (quotes from Feynman, 1974, p. 11). Misaligned incentives and pressure to publish positive results contribute to this situation (e.g., Smaldino & McElreath, 2016).

One of a kind? At the intersection of formal and empirical sciences. We believe that one of the main reasons for this is that ML stands, like few other disciplines, at the interface between formal sciences and real-world applications. Because ML has strong foundations in formal sciences such as mathematics, (theoretical) computer science (CS), and mathematical statistics, many ML researchers are accustomed to reasoning mathematically about abstract objects – ML methods – using formal proofs. On the other hand, ML can also very much be considered a (software) engineering science, to create practical systems that can learn and improve their performance by interacting with their environment. Lastly, and especially concerning experimentation in ML, there exists an applied statistical perspective with a focus on thorough inductive reasoning. With its tradition in data analysis and design of experiments, it emphasizes the empirical aspects of ML research.

These different perspectives, with their specific objectives, methodology, and terminology, have their unique virtues, but they also have their blind spots. The formal science perspective aims at providing absolute certainty and deep insights through the definition of abstract concepts and mathematical proofs but is often not well suited to explain complex real-world phenomena, as these concepts and proofs very often have to be based on strongly simplifying assumptions. The engineering perspective brought us incredible application improvements, but at the same time, not all conducted experiments are optimally designed to generalize results beyond the specific application context (which is also often only implicitly defined), as the references provided at the beginning demonstrate.

A statistical perspective, which we adopt here, is very sensitive to such empirical issues – explaining/analyzing real-world phenomena and generalizing beyond a specific context (inductive reasoning) – and thus particularly suited to explain 1) why ML is faced with non-replicable research, and 2) how a more complete and nuanced understanding of empirical research in ML can help to overcome this situation. With **empirical ML** we thus mean in a broad sense the systematic investigation of ML algorithms, techniques, and conceptual questions through simulations, experimentation, and observation. It deals with real objects: implementations of algorithms – which are usually more complex than their theoretical counterparts (e.g., Kriegel et al., 2017) – running on physical computers; data gathered and produced/simulated in the real world; and their interplay. Rather than focusing solely on theoretical analysis and proofs, empirical research emphasizes practical evaluations using real-world and/or synthetic data. Empirical ML, as understood here, requires a mindset very different from engineering and formal sciences and a different approach to methodology to allow for the full incorporation of the uncertainties inherent in dealing with real-world entities in experiments.

In our view, the discussed literature, raising similar points, has two main shortcomings: 1) they address only specific aspects of the problem and do not provide a comprehensive picture; 2) there is a confusion of terminology. For example, Bouthillier et al. (2019) distinguish between *exploratory* and *empirical* research. Nakkiran & Belkin (2022) use the term *good experimental research* and contrast it in particular with *improving applications*. Sculley et al. (2018) talk about *empirical advancements* and *empirical analysis* that are not complemented by sufficient *empirical rigor*. And Drummond (2006) discusses ML as an *experimental science* hardly using the term *empirical* at all.

To overcome these issues, we gather opinions and (empirical) evidence scattered across the literature and different domains and try to develop a **comprehensive synthesis**. For example, similar problems have been discussed in bioinformatics for some time (e.g., Yousefi et al., 2010; Boulesteix, 2010). We also take into account literature from other, more distant fields facing related issues, such as psychology and medicine. We believe this comprehensive picture will allow for a broader and deeper understanding of the complexity of the situation, which may at first glance appear to be rather easy to solve, e.g., by more (rigorous) statistical machinery or more open research artifacts. It is our conviction that without this deeper understanding, a situation that has been warned about in vain for so long cannot be overcome.

2. The Status Quo of Empirical ML

Recent advances. It is important to emphasize that there have been encouraging first steps in terms of empirical ML research recently. This includes the newly created publication formats *Transactions on Machine Learning Research* (TMLR), *Journal of Data-centric Machine Learning Research* (DMLR), or the NeurIPS *Datasets and Benchmarks Track* launched in 2021. These venues explicitly include in their scope, e.g., "reproducibility studies of previously published results or claims" (TMLR, n.d.), "systematic analyses of existing systems on novel datasets or benchmarks that yield important new insight" (DMLR, n.d.), and "systematic analyses of existing systems on novel datasets yielding important new insight" (NeurIPS, n.d.). Further examples are the *I Can't Believe It's Not Better!* (ICBINB) workshop series (e.g., Forde et al., 2020) and *Repository of Unexpected Negative Results* (ICBINB Initiative, n.d.) and efforts towards preregistration (e.g., Albanie et al., 2021) and reproducibility (e.g., Sinha et al., 2023). These developments, while very important, are not sufficient in our view to overcome the problems empirical ML faces. For example, while *computational reproducibility** may be a necessary condition, it is not a sufficient condition for replicability (e.g., Bouthillier et al., 2019). Furthermore, while the topics in the above formats cover many important aspects of empirical ML, we feel that they do not emphasize enough the importance of true replication of research, which is paramount from an empirical perspective. Most importantly, a situation in which a line of research warning us for a long time has been largely neglected will not be overcome by such practical changes alone. It also requires a change in awareness – of the importance of proper empirical ML but maybe even more of its limitations; and that there are different, equally valid types of proper empirical inquiry. We see this lack of awareness evidenced by TMLR (n.d.) itself: "TMLR emphasizes technical correctness over subjective significance, to ensure that we facilitate scientific discourse on topics that may *not yet be accepted in mainstream venues* [emphasis added] but may be important in the future." This is expressed in the talk introducing TMLR, too.[3] Judging by the example of other empirical sciences, this general lack of awareness of proper empirical ML is certainly the most difficult thing to overcome. Below we discuss problems we identified as symptoms of this lack.

Problem 1: Lack of unbiased experiments and scrutiny. Most method comparisons are carried out as part of a paper introducing a new method and are usually biased in favor of the new method (see Section 1 for examples).
Sculley et al. (2018, p. 1) found that "[l]ooking over papers from the last year, there seems to be a clear trend of multiple groups finding that prior work in fast moving fields may have missed improvements or key insights due to things as simple as hyperparameter tuning studies[*] or ablation studies." Moreover, for a neutral method comparison study of survival prediction methods, it has been shown that method rankings can vary considerably depending on design and analysis choices made at the meta-level (e.g., the selected set of datasets, performance metric, aggregation method) and that any method – even a simple baseline – can achieve almost any rank (Nießl et al., 2022; see also Sonabend et al., 2022). We are convinced that it is not far-fetched to conclude that quite often results demonstrating the superiority of a newly proposed method are obtained by an experimental design favorable to that method.
As in other disciplines (Munafò et al., 2017), there are structural issues (e.g., publication bias, pressure to publish, lack of replication studies) and questionable practices (e.g., hypothesizing after the results are known [HARKing, Kerr, 1998] and p-hacking [Simonsohn et al., 2014]) that contribute to this lack of unbiased experiments and scrutiny. At the individual level, in particular, there is a lack of awareness that method comparisons performed as part of a paper introducing a new method are not well suited to draw reliable conclusions about a method beyond the datasets considered, especially if 1) the number of datasets considered is small (Dehghani et al., 2021; Koch et al., 2021), 2) there is meta-level overfitting on a single benchmark design (Recht et al., 2019; Beyer et al., 2020), 3) the set of datasets selected for the experiments is biased in favor of the newly proposed method, and 4) the authors are much more familiar with the new method than with its competitors, as is the case frequently (Johnson, 2002; Boulesteix et al., 2013; 2017). Furthermore, it is very easy to artificially make a method appear superior (e.g., Jelizarow et al., 2010; Norel et al., 2011; Nießl et al., 2022; Ullmann et al., 2023; Pawel et al., 2024; Nießl et al., 2024), and publication bias towards positive results is a strong incentive to engage in SOTA-hacking and demonstrate the superiority of a newly proposed method (Sculley et al., 2018; Gencoglu et al., 2019).

[3] TMLR - A New Open Journal For Machine Learning: https://youtu.be/Uc1r1LfJtds

At the system level there is a publication bias and a lack of replication and neutral method comparison studies (e.g., Boulesteix et al., 2013; 2015b). Sculley et al. (2018, p. 1) "observe that the rate of empirical advancement [larger and more complex experiments] may not have been matched by consistent increase in the level of empirical rigor across the field as a whole." In unsupervised learning, the problem is more pronounced than in supervised learning because "there is much less of a benchmarking tradition in the clustering area than in the field of supervised learning" (Van Mechelen et al., 2023, p. 2; see also Zimmermann, 2020).

Problem 2: Lack of legitimacy. The second problem highlights a specific aspect of the lack of awareness of how different types of empirical research can contribute to ML. The problem was addressed by Nakkiran & Belkin (2022) and we completely agree with their description:

"In mainstream ML venues, there is a perceived lack of legitimacy and a real lack of community for good experimental science – which neither proves a theorem nor improves an application. This effectively suppresses a mode of scientific inquiry which has historically been critical to scientific progress, and which has shown promise in both ML and in CS more generally" (Nakkiran & Belkin, 2022, p. 2).

They identify a strong bias of the ML community towards mathematical proofs (formal science perspective) and application improvements (engineering perspective), while *good experimental science* that does not focus on one of the above is not incentivized nor encouraged. Nakkiran & Belkin (2022) see this evidenced by the lack of specific subject areas, the exclusion from recent calls for papers, the lack of explicit guidelines for reviewers, and the organization of separate workshops on experimental scientific investigation at major ML conferences. In particular, reviewers "often ask for application improvements" and "for 'theoretical justification' for purely experimental papers" (Nakkiran & Belkin, 2022, pp. 2–3). Together these factors point to a structural problem hindering the recognition and promotion of some sorts of experimental research in ML.

We completely agree with this view but think it may not immediately be clear what distinguishes improving an application from good experimental science at first sight.[4] As we understand it, the focus on application improvement means that much empirical/experimental research in ML focuses on developing a new method and demonstrating that it is superior to existing methods by improving some (predictive) performance metric on specific real-world benchmark datasets. Good experimental science, on the other hand, is not about improving performance. It is about improving understanding and knowledge of a problem, a (class of) methods, or a phenomenon. Sculley et al. (2018, p. 2) emphasize that "[e]mpirical studies [in ML] have become challenges to be 'won', rather than a process for developing insight and understanding. Ideally, the benefit of working with real data is to tune and examine the behavior of an algorithm under various sampling distributions, to learn about the strengths and weaknesses of the algorithms, as one would do in controlled studies." And Rendsburg et al. (2020, p. 9) argue, "it is particularly important that our community actively attempts to understand the inherent inductive biases, strengths, and also the weaknesses of algorithms. Finding examples where an algorithm works is important – but maybe even more important is to understand under which circumstances the algorithm produces misleading results."

Problem 3: Lack of conceptual clarity and operationalization. There is a perceived lack of clarity about some important abstract concepts that are the objects of ML research on the one side and a lack of clear operationalization* in empirical investigations on the other side. Both aspects affect the validity of experiments in empirical ML.

This problem is the most complex one and probably for that reason the most difficult to describe in precise terms (cf. Saitta & Neri, 1998). However, since we think that this problem affects the validity of empirical research in ML in a fundamental way, an account of empirical ML that does not attempt to make it tangible would be incomplete. We aim to narrow down the problem by explicating examples for supervised learning and unsupervised learning.

In other sciences such as psychology and physics, *validity**, the fact that the experimental measurement process actually measures what it is intended to measure, is fundamental. It inevitably depends on a strict and thorough operationalization in what way abstract concepts that are to be measured relate to measurable entities in the real world. Note that "[o]perational analysis is an excellent diagnostic tool for revealing where our knowledge is weak, in order to guide our efforts to strengthening it. The Bridgmanian ideal[*] is always to back up concepts with operational definitions, that is, to ensure that every concept is independently measurable in every circumstance under which it is used" (Chang, 2004, p. 147). It is puzzling that validity and other quality criteria of empirical research have gained little attention in ML so far (e.g., Myrtveit et al., 2005; Segebarth et al., 2020; Raji et al., 2021).

Experimental validity in supervised learning. For supervised learning, the problem can be exemplified by the question of inference from experimental results on real data in method comparison and evaluation studies.[5] Typically, the goal is to generalize the observed performance difference between methods to datasets that were not included in a study, which would require specifying when datasets are from the same/different domain. The problem is that it is not

[4] To avoid misunderstandings: we do consider mathematical proofs and application improvements very valuable research!

[5] Another example independently affecting validity is underspecification, which "is common in modern ML pipelines, such as those based on deep learning" (D'Amour et al., 2022, p. 2).

at all clear in what sense results obtained from one set of real datasets can be generalized to any other set of datasets, as this would require a clear understanding of the distribution of the data-generating processes by which each dataset is generated (e.g., Aha, 1992; Salzberg, 1997; Boulesteix et al., 2015a; Herrmann, 2022; Strobl & Leisch, 2024). Without a definition of the population of data-generating processes, i.e., (some) clarity about an abstract concept, it can be argued that it is not clear what a real data comparison study actually measures. In other words, the collection of datasets considered "will not be representative of real data sets in any formal sense" (Hand, 2006, p. 12). Dietterich (1998, p. 4) even went so far as claiming that how to perform benchmark experiments on real datasets properly is "perhaps the most fundamental and difficult question in machine learning."

Experimental validity in unsupervised learning. Arguably, the situation is even more involved in unsupervised learning (e.g., Kleinberg, 2002; von Luxburg et al., 2012; Zimek & Filzmoser, 2018; Herrmann, 2022). First of all, "there is no [...] direct measure of success. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms" (Hastie et al., 2009, p. 487). This is aggravated by an ambiguity about the abstract concepts of interest. Consider, for example, cluster analysis.[6] Usually, clusters are conceptualized as the modes of a mixture of (normal) distributions. However, there is a different perspective that considers cluster analysis from a topological perspective and conceptualizes clusters as the connected components of a dataset (Niyogi et al., 2011). It is not clear if these different notions of clusters 1) conceptualize clusters equally well, 2) can be related to the same real-world entities, and 3) whether clustering methods developed based on these different notions are equally suitable for all clustering problems. There is some evidence that suggests this is not the case (Herrmann et al., 2023a).

Problem summary. We argue that much empirical ML research is prone to overly optimistic, unreliable, and difficult-to-refute judgments and conclusions. Many experiments in empirical ML research are based on insufficiently operationalized experimental setups, partially due to ambiguous and inconclusive conceptualizations underlying the experiments. To draw more reliable conclusions, we need more explicit, context-specific operationalizations and clearer delineations of the abstract concepts that are to be investigated. Recall that "[o]perational analysis is an excellent diagnostic tool for revealing where our knowledge is weak, in order to guide our efforts in strengthening it" (Chang, 2004, p. 147). That sometimes good experimental research is not encouraged enough in ML (see Problem 2) and biased experiments still occur more often than desirable (see Problem 1), exacerbates the situation considerably. The former is an excellent approach for improving insight and understanding in the sense outlined above, biased experiments tend to make this more difficult. These aspects are becoming specifically important in deep learning where the sheer complexity of today's models, especially of foundation models, makes mathematical analysis extremely difficult. Instead, the analysis often needs to be largely experimental and thus requires thorough experimentation at the highest possible level.

[6] See also Herrmann et al. (2023b) for outlier detection.

3. Improving the Status Quo: More Richness in Empirical Methodological Research

A unifying view: We need exploratory and confirmatory. Confirmatory research*, also known as hypothesis-testing research, aims to test preexisting hypotheses to confirm or refute existing theories. Researchers design specific studies to evaluate hypotheses derived from existing knowledge experimentally. Typically, this involves a structured and predefined research design, a priori hypotheses, and often statistical analyses to draw conclusive inferences. In contrast, exploratory research is an open-ended approach that aims to gain insight and understanding in a new or unexplored area. It is often conducted when little is known about the phenomenon under study. It involves gathering information, identifying patterns, and formulating specific hypotheses for further investigation. One of our main points is that to improve empirical ML towards more thorough, reliable, and insightful methodological research both exploratory and confirmatory research are needed in ML (cf. Tukey, 1980). In general, the problems described can be placed in this broader epistemic context. We argue that most empirical research in ML is perceived as confirmatory research, when it should rather be considered to be exploratory from an epistemic perspective (see also Bouthillier et al., 2019). At the same time, purely exploratory methodological research focusing on improving insight and understanding experimentally (cf. Dietterich, 1990) and research like neutral method comparison and replication studies, which can be considered more rigorous in the confirmatory sense, are not seen as an equally important contribution to the field.

For the time being, it is worth making this distinction, yet, we discuss why it is an oversimplification from an epistemic perspective in Section 4 – even more so, because we distinguish two types of exploratory empirical methodological research in the following:[7] *insight-oriented exploratory research** in contrast to *method-developing exploratory research**. We think insight-oriented exploratory research is what Nakkiran & Belkin (2022) mean by good experimental research, and what they mean by application improvements is a conflation of both method-developing exploratory and (supposedly) confirmatory research.

[7] It is not our intention to establish a precise terminology, but we think this structure will be of assistance to the reader.

More insight-oriented exploratory research. In principle, the good thing about moving towards more insight-oriented exploratory methodological research in ML is that there are no epistemological obstacles to overcome. The neighboring field *data mining and knowledge discovery* clearly has an exploratory nature and is very much in the spirit of Tukey's exploratory data analysis. There are also already some examples of influential ML research that can be considered insight-oriented and exploratory, e.g., Frankle & Carbin (2019), Belkin et al. (2019), Recht et al. (2019), Rendsburg et al. (2020), Zhang et al. (2021), or Power et al. (2021). So, rather than epistemic aspects, it is the incentives and attitudes in scientific practice towards this type of research that are an obstacle to its successful dissemination. In particular, an alleged lack of novelty and originality is often invoked, which leads to rejections. Yet, without the esteem expressed by acceptance for publication, in particular in major ML venues, there is simply little incentive to engage in exploratory ML research. More importantly, it reinforces the impression among students and young scientists that exploratory research is not an integral part of science. It is therefore necessary to stimulate, encourage, and provide opportunities to make such research visible. Nakkiran & Belkin (2022, pp. 4–5) propose to establish a special subject area within ML conferences for "Experimental Science of Machine Learning," focusing on "experimental investigation into the nature of learning and learning systems." The types of papers outlined include those with "surprising experiments," "empirical conjectures," "refining existing phenomena," "formalizing intuition," and presentation of "new measurement tool[s]," all aiming to improve the understanding of ML empirically. They also provide guidelines specifically tailored to the review of this type of research.

More (actual) confirmatory research. As outlined, we believe most current empirical ML research (i.e., application improvements) is a mixture of method-developing exploratory research and (supposedly) confirmatory research.[8] For this reason, we add a focus on well-designed, neutral method comparison and replication studies. The scrutiny and rigor these examples of (actual) confirmatory empirical research provide are sorely needed if we are to work toward more reliable and replicable research.

Neutral method comparison studies include experiments that are less biased in favor of newly proposed methods (Boulesteix et al., 2013; Lim et al., 2000; Ali & Smith, 2006; Fernández-Delgado et al., 2014). First, this includes prespecified, strictly adhered-to designs of the experimental setup, including in particular a clearly specified set of datasets and tasks. Ideally, neutral comparison studies focus on the comparison of *already existing* methods and are carried out by a group of authors approximately equally familiar with all the methods under consideration (Boulesteix et al., 2013). Such studies ensure more neutrality and are less prone to overly optimistic conclusions than studies proposing a method, since there is much less of an incentive to promote a particular method. Second, proper **uncertainty quantification** is required when analyzing empirical results in ML, especially w.r.t. the different stages of inference (e.g., model fitting, model selection, pipeline construction, and performance estimation) (see Nadeau & Bengio, 2003; Bengio & Grandvalet, 2004; Hothorn et al., 2005; Bates et al., 2023). Moreover, if **statistical significance testing** is to be conducted to test for *statistically* significant performance differences across different real-world datasets, as described, e.g., by Demšar (2006), Eugster et al. (2012), Boulesteix et al. (2015a), or Eisinga et al. (2017), the methodological rigor established in other empirical domains should be applied (Munafò et al., 2017), in particular, efforts towards prior sample size calculations are important (Boulesteix et al., 2017).

Moreover, we need more **replication studies** and **meta-studies**. These types of research face similar reservations as insight-oriented exploratory experimental research. However, replication studies are indispensable to assess the amount of non-replicable research and to prevent it from being increased. Such studies attempt to reach the same scientific conclusions as previous studies, to provide additional empirical evidence for observed phenomena. Meta-studies analyze and summarize the so accumulated evidence on a specific phenomenon. This process is the default to reach conclusions in other sciences and is important as single studies can be false and/or contradict each other. In ML, this can range from studies that attempt to replicate an experiment exactly (e.g., Lohmann et al., 2022) or slightly modify an experiment's design (e.g., by using a different set of data in the replication of a neutral comparison study) to more comprehensive tuning and ablation studies of experiments conducted in method-developing research (e.g., Rendsburg et al., 2020; Kobak & Linderman, 2021). The latter certainly overlaps with insight-oriented exploratory research. It is important to emphasize that it is in the nature of things that a replication is not an original or novel scientific contribution in the conventional sense, and not necessarily can important new insights be gained beyond the replication of previously observed results. Rather, it is an explicit attempt to arrive at the same results and conclusions as a previous study. The scientific relevance, which is well acknowledged in other empirical sciences such as physics or medicine, lies in gathering additional empirical evidence for a hypothesis through a successful replication. Moreover, a replication study may, but does not necessarily, also raise epistemic questions, point to experimental improvements, or provide refined concepts, especially in failed replications.

[8] In a sense, this limits the potential of the former and renders the latter largely useless, with biased experiments as a result.

More infrastructure. To achieve this, practical limitations also need to be overcome. We require more dedicated infrastructure to make the proposed forms of research more (easily) realizable. In particular, there is a need for more and better open databases of well-curated and well-understood datasets such as OpenML (Vanschoren et al., 2013) or OpenML Benchmarking Suites (Bischl et al., 2021). Moreover, well-maintained open-source software for systematic benchmark experiments, such as the AutoML Benchmark (Gijsbers et al., 2024), HPOBench (Eggensperger et al., 2021), NAS-Bench-Suite (Mehta et al., 2022), or AlgoPerf (Dahl et al., 2023), are needed. Platforms for public leaderboards and model sharing (e.g., Hugging Face) are another important aspect, although some of these platforms are geared towards horse racing based on predictive performance and therefore do not necessarily also provide scientific insights or interpretability. Yet, the standards and automatic nature of such platforms have the advantage that they offer concrete reference points for criticism and debate. Finally, reviewer guidelines implementing our suggestions and dedicated venues for currently hard-to-publish empirical work will allow the full potential of empirical ML to be realized (Sculley et al., 2018; Nakkiran & Belkin, 2022).
Moreover, without **more education**, none of this will be possible. Given the different perspectives – formal science, engineering, statistical – from which ML can be viewed, it is very difficult to include each in the appropriate depth in a single study program. While a recent survey of 101 undergraduate data science programs in the U.S. showed that all included an introductory course in statistics (Bile Hassan et al., 2021), statistics has only recently (2023) been included as a core topic in the curriculum recommendations for CS* (Joint Task Force on Computing Curricula, 2023). It is also questionable if introductory courses are sufficient to avoid crucial gaps that can lead to the adoption of questionable research practices (cf. Gigerenzer, 2018). Furthermore, nearly no study program contains a dedicated course on *design and analysis of (computer) experiments* (Santner et al., 2003; Box et al., 2005; Dean et al., 2017), which we deem especially relevant for our context here. In general, we agree with De Veaux et al. (2017, pp. 16–17) that many "courses traditionally found in computer science, statistics, and mathematics offerings should be redesigned for the data science [or ML] major in the interest of efficiency and the potential synergy that integrated courses would offer."

Finally, we would like to offer **concrete and practicable advice** to specific target groups, in addition to the general recommendations above.
Advice for junior researchers. (1) Read the positive examples of insight-oriented exploratory research in ML (listed above), about the design of experiments, the critical discussion on statistical testing, and the basics of philosophy of science. (2) Educate yourself in Open Science practices (e.g., see The Turing Way Community, 2023). (3) Engage with researchers from other disciplines as data (the one on which ML models are trained) can only really be understood if one understands how it was generated. (4) Consider making empirical research in ML a (partial) research focus.
Advice for senior researchers. (1) Allow your junior researchers to write (great) papers on empirical aspects of ML, even if those may be relatively difficult to publish in major venues for now. Our personal experience is that these papers can still be highly cited and become very influential. (2) Learn from other fields; what we are experiencing in terms of non-replicable research is not a new phenomenon. (3) Please do not perceive this paper as an attack on ML but rather as an honest attempt to improve it and, more importantly, to improve its impact.
Advice for venue organizers and editors (see also Nakkiran & Belkin, 2022). (1) Encourage all forms of proper empirical ML to be submitted (in particular, this includes insight-oriented exploratory research), e.g., by creating special tracks or adding keywords but also by allowing such work on main tracks. The idea is to create special measures for the topic to increase awareness but not to isolate or ban all such papers to special (workshop) tracks with (potentially) lower perceived value. (2) Consider giving out awards for positive examples of these types of research. (3) Consider establishing positions like *reproducibility* and *replicability* editors for venues and journals. (4) Give concrete advice for best practices, so authors and reviewers have clear guidelines to follow. Note that this should not be confined to asking "Were the empirical results subjected to statistical tests?" (without further information); this is close to the opposite of what we think is needed.[9]

4. Beyond the Status Quo: Rethinking Empirical ML as a Maturing Science

The exploratory-confirmatory research continuum. With ML's strong foundation in formal sciences, where absolute certainty can be achieved by formal proofs, the clear distinction between exploratory and confirmatory research that has been invoked so far may seem natural. Yet, from an empirical perspective, i.e., whenever one deals with entities in the real world, it is itself an oversimplifying dichotomy, and empirical research is better thought of as a continuum from exploratory to confirmatory, with an ideal of purely exploratory research at one end and of strictly confirmatory research at the other (e.g., Wagenmakers et al., 2012; Oberauer & Lewandowsky, 2019; Szollosi & Donkin, 2021; Scheel et al., 2021; Devezer et al., 2021; Rubin & Donkin, 2022; Höfler et al., 2022; Fife & Rodgers, 2022).

[9] One anonymous reviewer also suggested that venues start collecting metadata on reasons for rejection. Such data could serve as a basis to evaluate if certain types of ML research face a systematic bias.

Based on that notion, Fife & Rodgers (2022) argue that "psychology may not be mature enough to justify confirmatory research" (p. 453) and that "[t]he maturity of any science puts a cap on the exploratory/confirmatory continuum" (p. 462). Given the similarities between research in psychology and ML as described by Hullman et al. (2022), we think similar holds for ML and we suggest ML should be considered as a maturing (empirical) science as well.[10] Hullman et al. (2022, p. 355) "identify common themes in reform discussions, like overreliance on asymptotic theory and non-credible beliefs about real-world data-generating processes." That said, confirmatory research in ML as advocated in Section 3 is still very different from *strict* confirmatory research in other disciplines. Rather it can be seen as *rough* confirmatory research (Fife & Rodgers, 2022; Tukey, 1973) that follows the same principles, but – as outlined – it is unclear how results can be generalized (e.g., using statistical tests) – a cornerstone of strict confirmatory research. But this should not be taken as a caveat. Rough confirmatory research allows for flexibility that strict confirmatory research does not (Fife & Rodgers, 2022).

The framework proposed by Heinze et al. (2024, p. 1) can be seen as a way of mapping this rather abstract idea into more concrete guidelines for scientific practice. In the context of biostatistics, they propose to consider four phases of methodological research, analogous to clinical research in drug development: "(I) proposing a new methodological idea while providing, for example, logical reasoning or proofs, (II) providing empirical evidence, first in a narrow target setting, then (III) in an extended range of settings and for various outcomes, accompanied by appropriate application examples, and (IV) investigations that establish a method as sufficiently well-understood to know when it is preferred over others and when it is not; that is, its pitfalls."

Statistical significance tests: Words of caution, revisited! The problem of empirical research as a continuum is more involved epistemologically and cannot be discussed in full detail here. An important aspect that needs to be discussed is its relation to the misguided use of statistical testing. This point has been made by Drummond (2006) before and in more detail. We revisit it here, enriching it with more recent literature on the issue. In particular, *routinely* adding statistical machinery to an (already underspecified and/or biased) experimental design to test for *statistically significant* differences in performance – as is frequently done and/or explicitly asked for (e.g., Henderson et al., 2018; Marie et al., 2021) – does not improve the epistemic relevance of the results by much nor does it add much additional insight over other data aggregations. In fact, "[s]tatistical significance was never meant to imply scientific importance," and you should not "conclude anything about scientific or practical importance based on statistical significance (or lack thereof)" (Wasserstein et al., 2019, pp. 2, 1). On the contrary, the misguided beliefs in and use of *statistical rituals* (Gigerenzer, 2018) is largely responsible for the replication crisis in other empirical disciplines.

The reasons are complex. First of all, the modern theory of statistical hypothesis testing (SHT) is a conflation of *two historically distinct types of testing theory**. Important epistemological questions about when statistical tests are appropriate are obscured by this mixed theory (e.g., Schneider, 2015; Gigerenzer & Marewski, 2015; Rubin, 2020). More importantly, specifically for experiments in ML, both theories are developed for experimental designs based on samples randomly drawn from a population of interest (Schneider, 2015). In general, the assumptions underlying the theory of statistical testing as an inferential tool are usually not met in many applications (Greenland, 2023). In fact, the editors of the *The American Statistician* special issue "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" went so far as to conclude, "based on [their] review of the articles in this special issue and the broader literature, that it is time to stop using the term 'statistically significant' entirely" (Wasserstein et al., 2019, p. 2).

Note that we want to warn against an *overemphasis* on as well as an *uncritical* use of statistical tests; we do not argue against statistical testing in general. Quite the contrary, we argue for a more diverse set of analysis tools (applied with care and critical reflection), including but not limited to statistical testing. We also want to stress that statistical testing cannot remedy more fundamental problems such as poor experimental design. To summarize the main points, we emphasize:

- Valid statistical testing inevitably depends on a thorough and well-designed experimental setup.
- Statistical testing should not be applied routinely and requires thought and careful preparation to be valid and insightful.
- Improper statistical testing and/or its uneducated interpretation are – widely acknowledged – a main driver for non-replicable results in other empirical sciences.
- The discussion about these issues has been going on for decades and has resulted in a large body of literature, some of which is condensed in the mentioned special issue of *The American Statistician*.

So, while we argue for more experiments in a confirmatory spirit to improve the status quo of empirical ML (see Section 3), especially using neutral method comparison and replication studies, we also emphasize that it is important to keep in mind their current epistemic limitations. In particular, we warn against common misconceptions about and inap-

[10] There are also differences between ML and psychology considerably simplifying our lives: we usually do not experiment on humans but algorithms on computers and have more control over experiments, larger sample sizes, and lower experimental costs.

propriate use of SHT. The problem is that the underlying "misunderstandings stem from a set of interrelated cognitive biases that reflect innate human compulsions which even the most advanced mathematical training seems to do nothing to staunch, and may even aggravate: Dichotomania, the tendency to reduce quantitative scales to dichotomies; nullism, the tendency to believe or at least act as if an unrefuted null hypothesis is true; and statistical reification, the tendency to forget that mathematical arguments say nothing about reality except to the extent the assumptions they make (which are often implicit) can be mapped into reality in a way that makes them all correct simultaneously" (Greenland, 2023, p. 911).

Most current empirical ML research should rather be viewed as exploratory. As outlined, confirmatory research aims to test preexisting hypotheses, while exploratory research involves gathering information, identifying patterns, and formulating specific hypotheses for further investigation. We think, currently, most of the empirical research in ML is conducted as part of a paper introducing a new method and is *fashioned* as confirmatory research even though it is exploratory in nature. In our view, this is reflected especially in the routine use of statistical tests to aggregate benchmark results: the exploratory phase of method development (e.g., trying out different method variants) largely invalidates post hoc statistical tests. As Strobl & Leisch (2024, p. 2) put it: "In methodological research, comparison studies are often published either with the explicit or implicit aim to promote a new method by means of showing that it outperforms existing methods." In other words, the conducted experiments are set up to *confirm* the (implicit) hypothesis that the proposed method constitutes an improvement. Systemic pressures and conventions, as well as ML's strong roots in formal sciences and focus on improving applications, encourage this mindset and the practice of invoking confirmatory arguments. This is expressed in statements such as that "[i]t is well-known that reviewers ask for application improvements" and "for 'theoretical justification' for purely experimental papers, even when the experiments alone constitute a valid scientific contribution" (Nakkiran & Belkin, 2022, pp. 2–3). The problem with not emphasizing the exploratory nature is that "exploratory findings have a slippery way of 'transforming' into planned findings as the research process progresses" (Calin-Jageman & Cumming, 2019, p. 275) and "[a]t the bottom of that slippery slope one often finds results that don't reproduce" (Wasserstein et al., 2019, p. 3). Shifting the focus to an exploratory notion of method development is an opportunity to fully allow "to understand under which circumstances the algorithm produces misleading results" (Rendsburg et al., 2020, p. 9) and to "learn about [its] strengths and weaknesses" (Sculley et al., 2018, p. 2) and clearly report them.

5. Conclusion

This work offers perspectives on ML that outline how it should move from a field being largely driven by mathematical proofs and application improvements to *also* becoming a full-fledged empirical field driven by multiple types of experimental research. By providing concrete practical guidance but at the same time moderating expectations of what empirical research can achieve, we wish to contribute to greater overall reliability and trustworthiness.

For every don't, there is a do. However, we are aware that our explanations may initially leave the reader unsatisfied when it comes to translating the conclusions into scientific practice. For example, those who were hoping for guidelines on the correct use of statistical tests may well be at a complete loss. However, we do not believe that this is actually the case. If you are inclined to perform statistical tests as described by Demšar (2006), do so, but also be aware of the Do-lists described by Wasserstein et al. (2019, Ch. 3, 7). In this regard, we consider the following comment by Wasserstein et al. (ib., p. 6) very noteworthy: "Researchers of any ilk may rarely advertise their personal modesty. Yet, the most successful ones cultivate a practice of being modest throughout their research, by understanding and clearly expressing the limitations of their work." Furthermore, do not only rely on real data, use simulated data as well. Simulations are an excellent tool for operationalization, i.e., mapping abstract concepts to measurable entities. Yet, the most important point is that we should be open to different ways of doing experimental research and should not penalize research just because it does not follow certain established conventions. As Nakkiran & Belkin (2022, p. 6) put it: "Each paper must be evaluated on an individual basis"; this is challenging, but they suggest guidelines. The community should take them up to address this issue.

Embracing inconclusiveness. Summarizing the perspectives on empirical ML covered here, and returning to the idea of mature sciences, we believe that for ML to mature as an (empirical) science, a greater awareness of some epistemic limitations, but also of the plurality of ways to gain insights, might be all it needs. We believe that if empirical research is one thing, it is not conclusive and no single empirical study can prove anything with absolute certainty. It must be scrutinized, repeated, and reassessed in a sense of *epistemic iteration** (Chang, 2004). That said, we conclude by quoting Chang's thoughts (ib., p. 243) on science in general: "If something is actually uncertain, our knowledge is superior if it is accompanied by an appropriate degree of doubt rather than blind faith. If the reasons we have for a certain belief are inconclusive, being aware of the inconclusiveness prepares us better for the possibility that other reasons may emerge to overturn our belief. With a critical awareness of uncertainty and inconclusiveness, our knowledge reaches a higher level of flexibility and sophistication."

Impact Statement

This position paper aims to advance machine learning by addressing practical challenges and epistemic constraints of empirical research that are often overlooked. We believe this has implications for machine learning research in general, as it can help to improve the reliability and credibility of research results. We also believe that our contribution can have a broader positive social and ethical impact by preventing misdirected efforts and resources.

Acknowledgments

We thank the four anonymous reviewers for their valuable comments and suggestions. Katharina Eggensperger is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Anne-Laure Boulesteix was partly funded by DFG grant BO3139/9-1.

References

Aha, D. W. Generalizing from case studies: A case study. In Sleeman, D. and Edwards, P. (eds.), *Machine Learning Proceedings 1992*, pp. 1–10, San Francisco, CA, United States, 1992. Morgan Kaufmann. doi:10.1016/B978-1-55860-247-2.50006-1.

Albanie, S., Henriques, J., Bertinetto, L., Hernandez-Garcia, A., Doughty, H., and Varol, G. The pre-registration workshop: An alternative publication model for machine learning research [Workshop]. *Thirty-Fifth Conference on Neural Information Processing Systems*, Online, 2021. https://neurips.cc/Conferences/2021/Schedule?showEvent=21885.

Ali, S. and Smith, K. A. On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138, 2006. doi:10.1016/j.asoc.2004.12.002.

Barba, L. A. Terminologies for reproducible research. *arXiv:1802.03311 [cs.DL]*, 2018. doi:10.48550/arXiv.1802.03311.

Bates, S., Hastie, T., and Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, pp. 1–12, 2023. doi:10.1080/01621459.2023.2197686.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi:10.1073/pnas.1903070116.

Bengio, Y. and Grandvalet, Y. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004. https://www.jmlr.org/papers/v5/grandvalet04a.html.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24, Granada, Spain, 2011. Curran Associates, Inc. https://papers.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Bergstra, J., Yamins, D., and Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, pp. 115–123, Atlanta, GA, United States, 2013. PMLR. https://proceedings.mlr.press/v28/bergstra13.html.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. Are we done with ImageNet? *arXiv:2006.07159 [cs.CV]*, 2020. doi:10.48550/arXiv.2006.07159.

Bile Hassan, I., Ghanem, T., Jacobson, D., Jin, S., Johnson, K., Sulieman, D., and Wei, W. Data science curriculum design: A case study. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pp. 529–534, Online, 2021. Association for Computing Machinery. doi:10.1145/3408877.3432443.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Gomes Mantovani, R., van Rijn, J., and Vanschoren, J. OpenML benchmarking suites. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Online, 2021. Curran Associates, Inc. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/c7e1249ffc03eb9ded908c236bd1996d-Abstract-round2.html.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., and Lindauer, M. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2): e1484, 2023. doi:10.1002/widm.1484.

Boulesteix, A.-L. Over-optimism in bioinformatics research. *Bioinformatics*, 26(3):437–439, 2010. doi:10.1093/bioinformatics/btp648.

Boulesteix, A.-L., Lauer, S., and Eugster, M. J. A. A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562, 2013. doi:10.1371/journal.pone.0061562.

Boulesteix, A.-L., Hable, R., Lauer, S., and Eugster, M. J. A. A statistical framework for hypothesis testing in real data comparison studies. *The American Statistician*, 69(3): 201–212, 2015a. doi:10.1080/00031305.2015.1005128.

Boulesteix, A.-L., Stierle, V., and Hapfelmeier, A. Publication bias in methodological computational research. *Cancer Informatics*, 14(S5):11–19, 2015b. doi:10.4137/CIN.S30747.

Boulesteix, A.-L., Wilson, R., and Hapfelmeier, A. Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, 17 (1):138, 2017. doi:10.1186/s12874-017-0417-2.

Bouthillier, X., Laurent, C., and Vincent, P. Unreproducible research is reproducible. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, pp. 725–734, Long Beach, CA, United States, 2019. PMLR. https://proceedings.mlr.press/v97/bouthillier19a.html.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P. Accounting for variance in machine learning benchmarks. In Smola, A., Dimakis, A., and Stoica, I. (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, Online, 2021. https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley Series in Probability and Statistics. John Wiley Wiley and Sons, 2nd edition, 2005.

Bridgman, P. W. *The Logic of Modern Physics*. Macmillan, 1927.

Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., and Boulesteix, A.-L. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, 22(1):152, 2021. doi:10.1186/s13059-021-02365-4.

Calin-Jageman, R. J. and Cumming, G. The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, 73(sup1): 271–280, 2019. doi:10.1080/00031305.2018.1518266.

Campell, D. T. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4): 297–312, 1957. doi:10.1037/h0040950.

Chang, H. *Inventing temperature: Measurement and scientific progress*. Oxford University Press, 2004.

Chang, H. Operationalism. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021. https://plato.stanford.edu/archives/fall2021/entries/operationalism/.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. doi:10.1016/j.jclinepi.2019.02.004.

Conference on Neural Information Processing Systems. NeurIPS 2023 Datasets and Benchmarks Track, n.d. Retrieved January 31, 2024, from https://neurips.cc/Conferences/2023/CallForDatasetsBenchmarks.

Dahl, G. E., Schneider, F., Nado, Z., Agarwal, N., Sastry, C. S., Hennig, P., Medapati, S., Eschenhagen, R., Kasimbeg, P., Suo, D., Bae, J., Gilmer, J., Peirson, A. L., Khan, B., Anil, R., Rabbat, M., Krishnan, S., Snider, D., Amid, E., Chen, K., Maddison, C. J., Vasudev, R., Badura, M., Garg, A., and Mattson, P. Benchmarking neural network training algorithms. *arXiv:2306.07179 [cs.LG]*, 2023. doi:10.48550/arXiv.2306.07179.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23:1–61, 2022. https://www.jmlr.org/papers/v23/20-1335.html.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruviluamala, N., Uhlig, P. X., Washington, T. M., Wesley, C. L., White, D., and Ye, P. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4:15–30, 2017. doi:10.1146/annurev-statistics-060116-053930.

Dean, A., Voss, D., and Draguljić, D. *Design and Analysis of Experiments*. Springer Texts in Statistics. Springer, 2nd edition, 2017. doi:10.1007/978-3-319-52250-0.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery. *arXiv:2107.07002 [cs.LG]*, 2021. doi:10.48550/arXiv.2107.07002.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. https://jmlr.org/papers/v7/demsar06a.html.

Devezer, B., Navarro, D. J., Vandekerckhove, J., and Buzbas, E. O. The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3):200805, 2021. doi:10.1098/rsos.200805.

Dhiman, P., Ma, J., Andaur Navarro, C. L., Speich, B., Bullock, G., Damen, J. A. A., Hooft, L., Kirtley, S., Riley, R. D., Van Calster, B., Moons, K. G. M., and Collins, G. S. Risk of bias of prognostic models developed using machine learning: A systematic review in oncology. *Diagnostic and Prognostic Research*, 6:13, 2022. doi:10.1186/s41512-022-00126-w.

Dietterich, T. G. Exploratory research in machine learning. *Machine Learning*, 5(1):5–9, 1990. doi:10.1007/BF00115892.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi:10.1162/089976698300017197.

Drummond, C. Machine learning as an experimental science (revisited). In *AAAI Workshop on Evaluation Methods for Machine Learning*, Boston, MA, United States, 2006. https://aaai.org/papers/ws06-06-002-machine-learning-as-an-experimental-science-revisited/.

Drummond, C. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th Annual International Conference on Machine Learning*, Montreal, Canada, 2009. https://www.site.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf.

Drummond, C. and Japkowicz, N. Warning: Statistical benchmarking is addictive. Kicking the habit in machine learning. *Journal of Experimental & Theoretical Artificial Intelligence*, 22(1):67–80, 2010. doi:10.1080/09528130903010295.

Eggensperger, K., Müller, P., Mallik, N., Feurer, M., Sass, R., Klein, A., Awad, N., Lindauer, M., and Hutter, F. HPOBench: A collection of reproducible multi-fidelity benchmark problems for HPO. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Online, 2021. Curran Associates, Inc. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/93db85ed909c13838ff95ccfa94cebd9-Abstract-round2.html.

Eisinga, R., Heskes, T., Pelzer, B., and Te Grotenhuis, M. Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. *BMC Bioinformatics*, 18(1):68, 2017. doi:10.1186/s12859-017-1486-2.

Elor, Y. and Averbuch-Elor, H. To SMOTE, or not to SMOTE? *arXiv:2201.08528 [cs.LG]*, 2022. doi:10.48550/arXiv.2201.08528.

Eugster, M. J. A., Hothorn, T., and Leisch, F. Domain-based benchmark experiments: Exploratory and inferential analysis. *Austrian Journal of Statistics*, 41(1):5–26, 2012. doi:10.17713/ajs.v41i1.185.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014. https://jmlr.org/papers/v15/delgado14a.html.

Ferrari Dacrema, M., Boglio, S., Cremonesi, P., and Jannach, D. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems*, 39(2):1–49, 2021. doi:10.1145/3434185.

Feurer, M. and Hutter, F. Hyperparameter optimization. In Hutter, F., Kotthoff, L., and Vanschoren, J. (eds.), *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pp. 3–33. Springer, 2019. doi:10.1007/978-3-030-05318-5_1.

Feynman, R. P. Cargo cult science. *Engineering and Science*, 37(7):10–13, 1974. Transcript of commencement address given at the California Institute of Technology. Available at http://calteches.library.caltech.edu/51/2/CargoCult.htm.

Fife, D. A. and Rodgers, J. L. Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist*, 77(3):453–466, 2022. doi:10.1037/amp0000886.

Forde, J. Z., Ruiz, F., Pradier, M. F., and Schein, A. I can't believe it's not better! Bridging the gap between theory and empiricism in probabilistic machine learning [Workshop]. *Thirty-Fourth Conference on Neural Information Processing Systems*, Online, 2020. https://neurips.cc/virtual/2020/protected/workshop_16124.html.

Foster, C. Methodological pragmatism in educational research: From qualitative-quantitative to exploratory-confirmatory distinctions. *International Journal of Research & Method in Education*, 47(1):4–19, 2024. doi:10.1080/1743727X.2023.2210063.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations*, New Orleans, LA, United States, 2019. https://openreview.net/forum?id=rJl-b3RcF7.

Franklin, A. and Perovic, S. Experiment in physics. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023. https://plato.stanford.edu/archives/fall2023/entries/physics-experiment/.

Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., Leinonen, J., and Huttunen, H. HARK side of deep learning – From grad student descent to automated machine learning. *arXiv:1904.07633 [cs.LG]*, 2019. doi:10.48550/arXiv.1904.07633.

Gigerenzer, G. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018. doi:10.1177/2515245918771329.

Gigerenzer, G. and Marewski, J. N. Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41(2):421–440, 2015. doi:10.1177/0149206314547522.

Gijsbers, P., Bueno, M. L. P., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., and Vanschoren, J. AMLB: An AutoML benchmark. *Journal of Machine Learning Research*, 25:1–65, 2024. https://www.jmlr.org/papers/v25/22-0493.html.

Greenland, S. Connecting simple and precise *P*-values to complex and ambiguous realities (includes rejoinder to comments on "Divergence vs. decision *P*-values"). *Scandinavian Journal of Statistics*, 50(3):899–914, 2023. doi:10.1111/sjos.12645.

Gundersen, O. E. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197):20200210, 2021. doi:10.1098/rsta.2020.0210.

Hand, D. J. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006. doi:10.1214/088342306000000060.

Hastie, T., Tibshirani, R., and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 2009. doi:10.1007/978-0-387-84858-7.

Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., White, I. R., and Simulation Panel of the STRATOS initiative. Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal*, 66(1):2200222, 2024. doi:10.1002/bimj.202200222.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3207–3214, New Orleans, LA, United States, 2018. AAAI Press. doi:10.1609/aaai.v32i1.11694.

Herrmann, M. *Towards more reliable machine learning: Conceptual insights and practical approaches for unsupervised manifold learning and supervised benchmark studies*. PhD thesis, Ludwig-Maximilians-Universität München, Munich, Germany, 2022. doi:10.5282/edoc.30789.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):bbaa167, 2020. doi:10.1093/bib/bbaa167.

Herrmann, M., Kazempour, D., Scheipl, F., and Kröger, P. Enhancing cluster analysis via topological manifold learning. *Data Mining and Knowledge Discovery*, pp. 1–48, 2023a. doi:10.1007/s10618-023-00980-2.

Herrmann, M., Pfisterer, F., and Scheipl, F. A geometric framework for outlier detection in high-dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(3):e1491, 2023b. doi:10.1002/widm.1491.

Höfler, M., Scherbaum, S., Kanske, P., McDonald, B., and Miller, R. Means to valuable exploration: I. The blending of confirmation and exploration and how to resolve it. *Meta-Psychology*, 6, 2022. doi:10.15626/MP.2021.2837.

Hooker, J. N. Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1:33–42, 1995. doi:10.1007/BF02430364.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005. doi:10.1198/106186005X59630.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., and Narayanan, A. 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*, pp. 335–348, Oxford, United Kingdom, 2022. Association for Computing Machinery. doi:10.1145/3514094.3534196.

ICBINB Initiative. ICBINB Repository of Unexpected Negative Results, n.d. Retrieved January 31, 2024, from `http://icbinb.cc/icbinb-repository-of-unexpected-negative-results/`.

Jaeger, R. G. and Halliday, T. R. On confirmatory versus exploratory research. *Herpetologica*, 54:S64–S66, 1998. `https://www.jstor.org/stable/3893289`.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A.-L. Over-optimism in bioinformatics: An illustration. *Bioinformatics*, 26(16):1990–1998, 2010. doi:10.1093/bioinformatics/btq323.

Johnson, D. S. A theoretician's guide to the experimental analysis of algorithms. In Goldwasser, M. H., Johnson, D. S., and McGeoch, C. C. (eds.), *Data Structures, Near Neighbor Searches, and Methodology: Fifth and Sixth DIMACS Implementation Challenges*, volume 59 of *Series in Discrete Mathematics & Theoretical Computer Science*, pp. 215–250. American Mathematical Society, 2002.

Joint Task Force on Computing Curricula. *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. Association for Computing Machinery and IEEE Computer Society, 2013. doi:10.1145/2534860.

Joint Task Force on Computing Curricula. *Computer Science Curricula 2023 – The Final Report*. Association for Computing Machinery, IEEE Computer Society, and Association for the Advancement of Artificial Intelligence, 2023. The report is not yet listed on the ACM curricula recommendation website. The final version is available at `https://csed.acm.org/final-report/`.

Journal of Data-centric Machine Learning Research. Submission guidelines for authors, n.d. Retrieved January 31, 2024, from `https://data.mlr.press/submissions.html`.

Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4 (9):100804, 2023. doi:10.1016/j.patter.2023.100804.

Kerr, N. L. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3): 196–217, 1998. doi:10.1207/s15327957pspr0203_4.

Kimmelman, J., Mogil, J. S., and Dirnagl, U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, 12(5): e1001863, 2014. doi:10.1371/journal.pbio.1001863.

Kleinberg, J. An impossibility theorem for clustering. In Becker, S., Thrun, S., and Obermayer, K. (eds.), *Advances in Neural Information Processing Systems*, volume 15, Vancouver, Canada, 2002. MIT Press. `https://papers.neurips.cc/paper_files/paper/2002/hash/43e4e6a6f341e00671e123714de019a8-Abstract.html`.

Kobak, D. and Linderman, G. C. Initialization is critical for preserving global data structure in both *t*-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021. doi:10.1038/s41587-020-00809-z.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. Reduced, reused and recycled: The life of a dataset in machine learning research. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Online, 2021. Curran Associates, Inc. `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/3b8a614226a953a8cd9526fca6fe9ba5-Abstract-round2.html`.

Kriegel, H.-P., Schubert, E., and Zimek, A. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, 52(2):341–378, 2017. doi:10.1007/s10115-016-1004-2.

Langley, P. Machine learning as an experimental science. *Machine Learning*, 3(1):5–8, 1988. doi:10.1023/A:1022623814640.

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. Are we learning yet? A meta review of evaluation failures across machine learning. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Online, 2021. Curran Associates, Inc. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40(3):203–228, 2000. doi:10.1023/A:1007608224229.

Lindstrom, L. Cargo cults. In *The Open Encyclopedia of Anthropology*. Facsimile of the first edition in *The Cambridge Encyclopedia of Anthropology*, 2023. doi:10.29164/18cargo.

Lipton, Z. C. and Steinhardt, J. Troubling trends in machine learning scholarship. *arXiv:1807.03341 [stat.ML]*, 2018. doi:10.48550/arXiv.1807.03341.

Lohmann, A., Astivia, O. L. O., Morris, T. P., and Groenwold, R. H. H. It's time! Ten reasons to start replicating simulation studies. *Frontiers in Epidemiology*, 2:973470, 2022. doi:10.3389/fepid.2022.973470.

Lones, M. A. How to avoid machine learning pitfalls: A guide for academic researchers. *arXiv:2108.02497 [cs]*, 2023. doi:10.48550/arXiv.2108.02497.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are GANs created equal? A large-scale study. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, Montréal, Canada, 2018. Curran Associates Inc. https://papers.neurips.cc/paper_files/paper/2018/hash/e46de7e1bcaaced9a54f1e9d0d2f800d-Abstract.html.

Mannarswamy, S. and Roy, S. Evolving AI from research to real life – Some challenges and suggestions. In Lang, J. (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 5172–5179, Stockholm, Sweden, 2018. AAAI Press. doi:10.24963/ijcai.2018/717.

Marie, B., Fujita, A., and Rubino, R. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7297–7306, Online, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.566.

Mateus, P., Volmer, L., Wee, L., Aerts, H. J. W. L., Hoebers, F., Dekker, A., and Bermejo, I. Image based prognosis in head and neck cancer using convolutional neural networks: A case study in reproducibility and optimization. *Scientific Reports*, 13:18176, 2023. doi:10.1038/s41598-023-45486-5.

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Feuer, B., Hegde, C., Ramakrishnan, G., Goldblum, M., and White, C. When do neural nets outperform boosted trees on tabular data? In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, New Orleans, LA, United States, 2023. Curran Associates Inc. https://papers.neurips.cc/paper_files/paper/2023/hash/f06d5ebd4ff40b40dd97e30cee632123-Abstract-Datasets_and_Benchmarks.html.

McGeoch, C. C. Experimental analysis of algorithms. In Pardalos, P. M. and Romeijn, H. E. (eds.), *Handbook of Global Optimization: Volume 2*, pp. 489–513. Springer, 2002. doi:10.1007/978-1-4757-5362-2_14.

Mehta, Y., White, C., Zela, A., Krishnakumar, A., Zabergja, G., Moradian, S., Safari, M., Yu, K., and Hutter, F. NAS-Bench-Suite: NAS evaluation is (now) surprisingly easy. In *10th International Conference on Learning Representations*, Online, 2022. https://openreview.net/forum?id=0DLwqQLmqV.

Melis, G., Dyer, C., and Blunsom, P. On the state of the art of evaluation in neural language models. In *6th International Conference on Learning Representations*, Vancouver, Canada, 2018. https://openreview.net/forum?id=ByJHuTgA-.

Merriam-Webster. Epistemic. In *Merriam-Webster.com dictionary*, n.d. Retrieved May 1, 2024, from https://www.merriam-webster.com/dictionary/epistemic.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):1–9, 2017. doi:10.1038/s41562-016-0021.

Myrtveit, I., Stensrud, E., and Shepperd, M. Reliability and validity in comparative studies of software prediction models. *IEEE Transactions on Software Engineering*, 31 (5):380–391, 2005. doi:10.1109/TSE.2005.58.

Nadeau, C. and Bengio, Y. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi:10.1023/A:1024068626366.

Nakkiran, P. and Belkin, M. Incentivizing empirical science in machine learning. In *ML Evaluation Standards Workshop at ICLR 2022*, Online, 2022. https://ml-eval.github.io/assets/pdf/science_ml_proposal_2am.pdf.

Narang, S., Chung, H. W., Tay, Y., Fedus, W., Fevry, T., Matena, M., Malkan, K., Fiedel, N., Shazeer, N., Lan, Z., Zhou, Y., Li, W., Ding, N., Marcus, J., Roberts, A., and Raffel, C. Do transformer modifications transfer across implementations and applications? In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.465.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, 2019. doi:10.17226/25303.

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., and Boulesteix, A.-L. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2): e1441, 2022. doi:10.1002/widm.1441.

Nießl, C., Hoffmann, S., Ullmann, T., and Boulesteix, A.-L. Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment. *Biometrical Journal*, 66(1):2200238, 2024. doi:10.1002/bimj.202200238.

Nilsen, E. B., Bowler, D. E., and Linnell, J. D. C. Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology*, 57(4):842–847, 2020. doi:10.1111/1365-2664.13571.

Niyogi, P., Smale, S., and Weinberger, S. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663, 2011. doi:10.1137/090762932.

Norel, R., Rice, J. J., and Stolovitzky, G. The self-assessment trap: Can we all be better than average? *Molecular Systems Biology*, 7(1):537, 2011. doi:10.1038/msb.2011.70.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi:10.1073/pnas.1708274114.

Oberauer, K. and Lewandowsky, S. Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26 (5):1596–1618, 2019. doi:10.3758/s13423-019-01645-2.

Pawel, S., Kook, L., and Reeve, K. Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, 66(1): 2200091, 2024. doi:10.1002/bimj.202200091.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22:1–20, 2021. https://jmlr.org/papers/v22/20-303.html.

Plesser, H. E. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:76, 2018. doi:10.3389/fninf.2017.00076.

Popper, K. R. *The Logic of Scientific Discovery*. Routledge, 2002. The work was originally published in 1935 in German. The first English edition was published in 1959.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *Mathematical Reasoning in General Artificial Intelligence Workshop at ICLR 2021*, Online, 2021. https://mathai-iclr.github.io/papers/papers/MATHAI_29_paper.pdf.

Raff, E. A step toward quantifying independently reproducible machine learning research. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, Vancouver, Canada, 2019. Curran Associates, Inc. https://papers.neurips.cc/paper_files/paper/2019/hash/c429429bf1f2af051f2021dc92a8ebea-Abstract.html.

Raff, E. and Farris, A. L. A siren song of open source reproducibility. In *ML Evaluation Standards Workshop at ICLR 2022*, Online, 2022. doi:10.48550/arXiv.2204.04372.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the everything in the whole wide world benchmark. In Vanschoren, J. and Yeung, S.

(eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, Online, 2021. Curran Associates, Inc. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400, Long Beach, CA, United States, 2019. PMLR. https://proceedings.mlr.press/v97/recht19a.html.

Rendsburg, L., Heidrich, H., and von Luxburg, U. NetGAN without GAN: From random walks to low-rank approximations. In Daumé III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, pp. 8073–8082, Online, 2020. PMLR. https://proceedings.mlr.press/v119/rendsburg20a.html.

Riquelme, C., Tucker, G., and Snoek, J. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *6th International Conference on Learning Representations*, Vancouver, Canada, 2018. https://openreview.net/forum?id=SyYe6k-CW.

Roettger, T. B. Preregistration in experimental linguistics: Applications, challenges, and limitations. *Linguistics*, 59 (5):1227–1249, 2021. doi:10.1515/ling-2019-0048.

Rubin, M. "Repeated sampling from the same population?" A critique of Neyman and Pearson's responses to Fisher. *European Journal for Philosophy of Science*, 10:42, 2020. doi:10.1007/s13194-020-00309-6.

Rubin, M. and Donkin, C. Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology*, pp. 1–29, 2022. doi:10.1080/09515089.2022.2113771.

Saitta, L. and Neri, F. Learning in the "real world". *Machine Learning*, 30(2–3):133–163, 1998. doi:10.1023/A:1007448122119.

Salzberg, S. L. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–328, 1997. doi:10.1023/A:1009752403260.

Santner, T. J., Williams, B. J., and Notz, W. I. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer, 2003. doi:10.1007/978-1-4757-3799-8.

Scheel, A. M., Tiokhin, L., Isager, P. M., and Lakens, D. Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16 (4):744–755, 2021. doi:10.1177/1745691620966795.

Schneider, J. W. Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1):411–432, 2015. doi:10.1007/s11192-014-1251-5.

Schwab, S. and Held, L. Different worlds confirmatory versus exploratory research. *Significance*, 17(2):8–9, 2020. doi:10.1111/1740-9713.01369.

Sculley, D., Snoek, J., Wiltschko, A., and Rahimi, A. Winner's curse? On pace, progress, and empirical rigor. In *6th International Conference on Learning Representations – Workshop*, Vancouver, Canada, 2018. https://openreview.net/forum?id=rJWF0Fywf.

Segebarth, D., Griebel, M., Stein, N., von Collenberg, C. R., Martin, C., Fiedler, D., Comeras, L. B., Sah, A., Schoeffler, V., Lüffe, T., Dürr, A., Gupta, R., Sasi, M., Lillesaar, C., Lange, M. D., Tasan, R. O., Singewald, N., Pape, H.-C., Flath, C. M., and Blum, R. On the objectivity, reliability, and validity of deep learning enabled bioimage analyses. *eLife*, 9:e59780, 2020. doi:10.7554/eLife.59780.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547, 2014. doi:10.1037/a0033242.

Sinha, K., Forde, J. Z., Samiei, M., Ghosh, A., Sutawika, L., and Panigrahi, S. S. Announcing MLRC 2023. ML Reproducibility Challenge, 2023, October 18. Retrieved January 31, 2024, from https://reproml.org/blog/announcing_mlrc2023/.

Smaldino, P. E. and McElreath, R. The natural selection of bad science. *Royal Society Open Science*, 3(9):160384, 2016. doi:10.1098/rsos.160384.

Sonabend, R., Bender, A., and Vollmer, S. Avoiding c-hacking when evaluating survival distribution predictions with discrimination measures. *Bioinformatics*, 38(17): 4178–4184, 2022. doi:10.1093/bioinformatics/btac451.

Steup, M. Epistemology. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2006 edition, 2006. https://plato.stanford.edu/archives/spr2006/entries/epistemology/.

Steup, M. and Neta, R. Epistemology. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

https://plato.stanford.edu/archives/fall2020/entries/epistemology/.

Strobl, C. and Leisch, F. Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*, 66(1):2200104, 2024. doi:10.1002/bimj.202200104.

Szollosi, A. and Donkin, C. Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4):717–724, 2021. doi:10.1177/1745691620966796.

Tatman, R., VanderPlas, J., and Dane, S. A practical taxonomy of reproducibility for machine learning research. In *Reproducibility in Machine Learning Workshop at ICML 2018*, Stockholm, Sweden, 2018. https://openreview.net/forum?id=B1eYYK5QgX.

The Turing Way Community. *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Zenodo, 2023. doi:10.5281/zenodo.7625728.

Transactions on Machine Learning Research. Transactions on Machine Learning Research, n.d. Retrieved January 31, 2024, from https://jmlr.org/tmlr/index.html.

Transactions on Machine Learning Research. Submission guidelines and editorial policies, n.d. Retrieved January 31, 2024, from https://jmlr.org/tmlr/editorial-policies.html.

Trosten, D. J. Questionable practices in methodological deep learning research. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 4, 2023. doi:10.7557/18.6804.

Tukey, J. W. Exploratory data analysis as part of a larger whole. In *Proceedings of the Eighteenth Conference on the Design of Experiments in Army Research, Development and Testing*, pp. 1–10, Aberdeen, MD, United States, 1973. U.S. Army Research Office. https://apps.dtic.mil/sti/citations/AD0776910.

Tukey, J. W. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980. doi:10.2307/2682991.

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., and Boulesteix, A.-L. Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*, 17(1):211–238, 2023. doi:10.1007/s11634-022-00496-5.

van den Goorbergh, R., van Smeden, M., Timmerman, D., and Van Calster, B. The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 2022. doi:10.1093/jamia/ocac093.

Van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Hennig, C., Leisch, F., Steinley, D., and Warrens, M. J. A white paper on good research practices in benchmarking: The case of cluster analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(6): e1511, 2023. doi:10.1002/widm.1511.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi:10.1145/2641190.2641198.

von Luxburg, U., Williamson, R. C., and Guyon, I. Clustering: Science or art? In Guyon, I., Dror, G., Lemaire, V., Taylor, G., and Silver, D. (eds.), *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 65–79, Bellevue, WA, United States, 2012. PMLR. https://proceedings.mlr.press/v27/luxburg12a.html.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638, 2012. doi:10.1177/1745691612463078.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, 2019. doi:10.1080/00031305.2019.1583913.

Yousefi, M. R., Hua, J., Sima, C., and Dougherty, E. R. Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, 26(1):68–76, 2010. doi:10.1093/bioinformatics/btp605.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. doi:10.1145/3446776.

Zimek, A. and Filzmoser, P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6):e1280, 2018. doi:10.1002/widm.1280.

Zimmermann, A. Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1330, 2020. doi:10.1002/widm.1330.

Glossary

Bridgmanian ideal. Used by Chang (2004) to describe a specific notion of *operationalization*. Refers to Percy Williams Bridgman (1882–1961), Nobel laureate in physics for his work on high-pressure physics, who also made contributions to the philosophy of science. Operational analysis is the topic of his book *The Logic of Modern Physics*, in which he argues in particular that "[i]n general, we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*" (Bridgman, 1927, p. 5). This strict perspective on operationalization (also referred to as operationalism) has attracted a lot of criticism, see "Operationalism" in *The Stanford Encyclopedia of Philosophy* (Chang, 2021). In particular, Chang (2004, p. 148) points out that it builds on "an overly restrictive notion of meaning, which comes down to reduction of meaning to measurement, which [Chang] refer[s] to as *Bridgman's reductive doctrine of meaning*."

Cargo Cult Science. The term cargo cult refers to social movements that originated in Melanesia: "The modal cargo cult was an agitation or organised social movement of Melanesian villagers in pursuit of 'cargo' by means of renewed or invented ritual action that they hoped would induce ancestral spirits or other powerful beings to provide" (Lindstrom, 2023, p. 1). Richard Phillips Feynman (1918–1988), theoretical physicist and Nobel laureate, adapted the term to describe ritualized scientific practices which "follow all the apparent precepts and forms of scientific investigation, but [which are] missing something essential" (Feynman, 1974, p. 11).

Confirmatory research. Also known as hypothesis-testing research, aims to test preexisting hypotheses to confirm or refute existing theories. Researchers design specific studies to evaluate hypotheses derived from existing knowledge experimentally. Typically, this involves a structured and predefined research design, a priori hypotheses, and often statistical analyses to draw conclusive inferences. It is a well-established term in many fields other than ML. For example, general references are Schwab & Held (2020), Nosek et al. (2018), and Munafò et al. (2017). Field-specific references include Jaeger & Halliday (1998) or Nilsen et al. (2020) for biology, Wagenmakers et al. (2012) for psychology, Kimmelman et al. (2014) for preclinical research, Roettger (2021) for linguistics, or Foster (2024) for educational research. The term *confirmatory* might appear to be in conflict with the principle of falsification established by Popper (1959/2002). According to Popper, scientific theories cannot be conclusively confirmed, only falsified. It is important to emphasize that *confirmatory research* has a narrower scope rooted in Neyman-Pearson statistical testing theory (see the glossary entry on *Two historically distinct types of testing theory*). This theory provides a framework for a statistically justified decision between a null hypothesis and an alternative hypothesis based on the available data. The hypothesis to be established (e.g., there is an effect) is usually stated as the alternative hypothesis and confirmation means rejecting the null hypothesis (e.g., there is no effect) for the alternative.

Curricula recommendations for CS. The report *Computer Science Curricula 2013* lists "Intelligent Systems" (including basics in ML) as a Core (Tier2) topic but "still believe[s] it is not necessary for all CS programs to require a full course in probability theory [or statistics]" (Joint Task Force on Computing Curricula, 2013, p. 50). This has changed with the latest (2023) version insofar as statistics is now considered a CS Core topic in "Mathematical and Statistical Foundations", which is one of several knowledge areas (Joint Task Force on Computing Curricula, 2023).

Epistemic, epistemological. Both coming from the Greek word for knowledge or understanding, the terms are sometimes used synonymously and sometimes with distinct, more precise meanings. If the distinction is made, *epistemic* relates to knowledge itself, while *epistemological* relates to "the study of the nature and grounds of knowledge" (Merriam-Webster, n.d.), i.e., epistemology. For epistemology, an early edition of *The Stanford Encyclopedia of Philosophy* gives the following definition: "Defined narrowly, epistemology is the study of knowledge and justified belief. [...] Understood more broadly, epistemology is about issues having to do with the creation and dissemination of knowledge in particular areas of inquiry" (Steup, 2006). The most recent edition states in more abstract terms that "[m]uch recent work in formal epistemology is an attempt to understand how our degrees of confidence are rationally constrained by our evidence [...]" and that "epistemology seeks to understand one or another kind of *cognitive success* [...]" (Steup & Neta, 2020).

Epistemic iteration. Chang (2004) introduced the concept and defined it in his glossary as a "process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals. It differs crucially from mathematical iteration in that the latter is used to approach a correct answer that is known, or at least in principle knowable, by other means" (p. 253). For thorough discussions, see Chapters 1 (pp. 46–48) and 5.

Exploratory research. As also specified in the main body of the paper, refers to an open-ended approach that aims to gain insight and understanding in a new or unexplored area (in contrast to *confirmatory research*). It is often conducted when little is known about the phenomenon under study. It involves gathering information, identifying patterns, and formulating specific hypotheses for further investigation.

Hyperparameter tuning studies. Aim to find the best-performing configuration for an ML model class, including baselines (Feurer & Hutter, 2019; Bischl et al., 2023). Tuned models can then be compared more objectively and fairly. Hyperparameter tuning (or the lack of it) is an important source of variation in benchmark studies (Bouthillier et al., 2021) and has been shown to have a strong effect on the outcome of results (see for example the references in Bouthillier et al., 2021 or our introduction). Treating the hyperparameter optimization as part of the problem of quantifying the performance of an algorithm has been suggested by Bergstra et al. (2011) and Bergstra et al. (2013).

Insight-oriented exploratory research. Refers to experimental research in ML that aims to gain insight, rather than inventing/developing a new method. It does not necessarily involve a very specific hypothesis to be pursued, but it is about improving the understanding and knowledge of a problem, a (class of) existing methods, or a phenomenon.

Method-developing exploratory research. Refers to experimental research in ML carried out in the process of developing a new ML method. This can include method comparison experiments, but in particular, it refers to exploration that takes place during the development process. This may include, for example, trying different method variants or specifying hyperparameter configurations and implementation details.

Operationalization. Chang (2004, p. 256) provides the following definition in his glossary: "The process of giving operational meaning to a concept where there was none before. Operationalization may or may not involve the specification of explicit measurement methods." *Operational meaning* refers to "the meaning of a concept that is embodied in the physical operations whose description involves the concept." For a thorough discussion, see Chapter 4 (pp. 197–219).

Replicability (vs. reproducibility). There is no consistent use of these terms in the broader literature (for discussions, e.g., see Barba, 2018; Plesser, 2018; Gundersen, 2021; Pineau et al., 2021). We use the term reproducibility in a narrow technical sense (see the glossary entry on *computational reproducibility*). In contrast, replicability here means arriving at the same scientific conclusions in a broad sense. This terminology is in line with the National Academies of Sciences, Engineering, and Medicine (2019). In terms of the reliability of results, it means that replicability is more important than reproducibility. Note that Drummond (2009), for example, uses the terms the reverse way.

Reproducibility (computational). Means that the provided code technically achieves the same result on the provided data, and not that code, experimental design, or analysis are error-free and that we can qualitatively reach the same conclusions for the same general question under slightly different technical conditions. It is thus not a sufficient condition for replicability. Note that Tatman et al. (2018) differentiate three levels of reproducibility.

Two historically distinct types of testing theory. This refers to two approaches to statistical testing developed by Ronald Aylmer Fisher (1890–1962) on the one side and Jerzy Neyman (1894–1981) and Egon Sharpe Pearson (1895–1980) on the other. Only the former includes p-values and a single (null) hypothesis. The latter includes two hypotheses and hinges on statistical power and Type I and II errors (Schneider, 2015, p. 413). More generally, Fisher's approach is "[b]ased on the concept of a 'hypothetical infinite population'," "[r]oots in inductive philosophy" and "[a]pplies to any single experiment (short run)," while Neyman-Pearson's approach is "[b]ased on a clearly defined population," "[r]oots in deductive philosophy," and "[a]pplies only to ongoing, identical repetitions of an experiment, not to any single experiment (long run)" (Schneider, 2015, p. 415, Table 1).

Validity. Note that there is no concise definition of the term. In psychology, internal and external validity are differentiated in particular. According to Campell (1957, p. 297) internal validity asks if "in fact the experimental stimulus make some significant difference in this specific instance?" External validity, on the other hand, asks "to what populations, settings, and variables can this effect be generalized?" The former appears to be closely related to in-distribution generalization performance in ML, the latter to out-of-distribution generalization. In contrast, *The Stanford Encyclopedia of Philosophy* states for experiments in physics (Franklin & Perovic, 2023): "Physics, and natural science in general, is a reasonable enterprise based on *valid* [emphasis added] experimental evidence, criticism, and rational discussion." Several strategies that may be used to validate observations are specified. These include the following: 1) "Experimental checks and calibration, in which the experimental apparatus reproduces known phenomena"; 2) "Reproducing artifacts that are known in advance to be present"; 3) "Elimination of plausible sources of error and alternative explanations of the result"; 4) "Using the results themselves to argue for their validity"; 5) "Using an independently well-corroborated theory of the phenomena to explain the results"; 6) "Using an apparatus based on a well-corroborated theory"; 7) "Using statistical arguments." However, it is emphasized that "[t]here are many experiments in which these strategies are applied, but whose results are later shown to be incorrect [...]. Experiment is fallible. Neither are these strategies exclusive or exhaustive. No single one of them, or fixed combination of them, guarantees the validity of an experimental result" (Franklin & Perovic, 2023).