Troubling Trends in Machine-learning Scholarship

ZACHARY C. LIPTON AND
JACOB STEINHARDT

SOME ML PAPERS SUFFER FROM FLAWS THAT COULD MISLEAD THE PUBLIC AND STYMIE FUTURE RESEARCH.

Collectively, ML (machine learning) researchers are engaged in the creation and dissemination of knowledge about data-driven algorithms. In a given paper, researchers might aspire to any subset of the following goals, among others: to theoretically characterize what is learnable; to obtain understanding through empirically rigorous experiments; or to build a working system that has high predictive accuracy. While determining which knowledge warrants inquiry may be subjective, once the topic is fixed, papers are most valuable to the community when they act in service of the reader, creating foundational knowledge and communicating as clearly as possible.

What sorts of papers best serve their readers? Ideally, papers should accomplish the following: (1) provide intuition to aid the reader's understanding but clearly distinguish it from stronger conclusions supported by evidence; (2) describe empirical investigations that consider and rule out alternative hypotheses; (3) make clear the relationship between theoretical analysis and intuitive or empirical claims; and (4) use language

to empower the reader, choosing terminology to avoid misleading or unproven connotations, collisions with other definitions, or conflation with other related but distinct concepts.

Recent progress in machine learning comes despite frequent departures from these ideals. This installment of Research for Practice focuses on the following four patterns that appear to be trending in ML scholarship:

- Failure to distinguish between explanation and speculation.
- Failure to identify the sources of empirical gains (e.g., emphasizing unnecessary modifications to neural architectures when gains actually stem from hyperparameter tuning).
- "Mathiness"—the use of mathematics that obfuscates or impresses rather than clarifies (e.g., by confusing technical and nontechnical concepts).
- Misuse of language (e.g., by choosing terms of art with colloquial connotations or by overloading established technical terms).

While the causes behind these patterns are uncertain, possibilities include the rapid expansion of the community, the consequent thinness of the reviewer pool, and the often-misaligned incentives between scholarship and short-term measures of success (e.g., bibliometrics, attention, and entrepreneurial opportunity). While each pattern offers a corresponding remedy (don't do it), this article also makes suggestions on how the community might combat these troubling trends.

As the impact of machine learning widens, and the audience for research papers increasingly includes

students, journalists, and policy-makers, these considerations apply to this wider audience as well. By communicating more precise information with greater clarity, ML scholarship could accelerate the pace of research, reduce the on-boarding time for new researchers, and play a more constructive role in public discourse.

Flawed scholarship threatens to mislead the public and stymie future research by compromising ML's intellectual foundations. Indeed, many of these problems have recurred cyclically throughout the history of AI (artificial intelligence) and, more broadly, in scientific research. In 1976, Drew McDermott[26] chastised the AI community for abandoning self-discipline, warning prophetically that "if we can't criticize ourselves, someone else will save us the trouble." Similar discussions recurred throughout the 1980s, 1990s, and 2000s. In other fields, such as psychology, poor experimental standards have eroded trust in the discipline's authority.[33] The current strength of machine learning owes to a large body of rigorous research to date, both theoretical and empirical. By promoting clear scientific thinking and communication, our community can sustain the trust and investment it currently enjoys.

DISCLAIMERS

This article aims to instigate discussion, answering a call for papers from the ICML (International Conference on Machine Learning) Machine Learning Debates workshop. While we stand by the points represented here, we do not purport to offer a full or balanced viewpoint or to discuss

the overall quality of science in ML. In many aspects, such as reproducibility, the community has advanced standards far beyond what sufficed a decade ago.

Note that these arguments are made by *us*, against *us*—insiders offering a critical introspective look—not as sniping outsiders. The ills identified here are not specific to any individual or institution. We have fallen into these patterns ourselves, and likely will again in the future. Exhibiting one of these patterns doesn't make a paper *bad,* nor does it indict the paper's authors; however, all papers could be made stronger by avoiding these patterns.

While we provide concrete examples, our guiding principles are (1) to implicate ourselves; and (2) to select preferentially from the work of better-established researchers and institutions that we admire, to avoid singling out junior students for whom inclusion in this discussion might have consequences and who lack the opportunity to reply symmetrically. We are grateful to belong to a community that provides sufficient intellectual freedom to allow the expression of critical perspectives.

TROUBLING TRENDS

Each subsection that follows describes a trend; provides several examples (as well as positive examples that resist the trend); and explains the consequences. Pointing to weaknesses in individual papers can be a sensitive topic. To minimize this, the examples are short and specific.

Explanation vs. speculation

Research into new areas often involves exploration predicated on intuitions that have yet to coalesce into crisp

formal representations. Speculation is a way for authors to impart intuitions that may not yet withstand the full weight of scientific scrutiny. Papers often offer speculation in the guise of *explanations*, however, which are then interpreted as authoritative because of the trappings of a scientific paper and the presumed expertise of the authors.

For instance, in a 2015 paper, Sergey Ioffe and Christian Szegedy form an intuitive theory around a concept called *internal covariate shift*.[18] The exposition on internal covariate shift, starting from the abstract, appears to state technical facts. Key terms are not made crisp enough, however, to assume a truth value conclusively. For example, the paper states that batch normalization offers improvements by reducing changes in the distribution of hidden activations over the course of training. By which divergence measure is this change quantified? The paper never clarifies, and some work suggests that this explanation of batch normalization may be off the mark.[37] Nevertheless, the speculative explanation given by Ioffe and Szegedy has been repeated as fact—for example, in a 2015 paper by Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han,[31] which states, "It is well known that a deep neural network is very hard to optimize due to the internal-covariate-shift problem."

We have been equally guilty of speculation disguised as explanation. In a 2017 paper with Pang Wei Koh and Percy Liang,[42] I (Jacob Steinhardt) write that "the high dimensionality and abundance of irrelevant features... give the attacker more room to construct attacks," without conducting any experiments to measure the effect of dimensionality on attackability. In another paper with

Liang from 2015,[41] I (Steinhardt) introduce the intuitive notion of *coverage* without defining it, and use it as a form of explanation (e.g., "Recall that one symptom of a lack of coverage is poor estimates of uncertainty and the inability to generate high-precision predictions." Looking back, we desired to communicate insufficiently fleshed-out intuitions that were material to the work described in the paper and were reticent to label a core part of the argument as speculative.

In contrast to these examples, Nitish Srivastava et al.[39] separate speculation from fact. While this 2014 paper, which introduced dropout regularization, speculates at length on connections between dropout and sexual reproduction, a designated "Motivation" section clearly quarantines this discussion. This practice avoids confusing readers while allowing authors to express informal ideas.

In another positive example, Yoshua Bengio presents practical guidelines for training neural networks.[2] Here, the author carefully conveys uncertainty. Instead of presenting the guidelines as authoritative, the paper states: "Although such recommendations come... from years of experimentation and to some extent mathematical justification, they should be challenged. They constitute a good starting point... but very often have not been formally validated, leaving open many questions that can be answered either by theoretical analysis or by solid comparative experimental work."

Failure to identify the sources of empirical gains

The ML peer-review process places a premium on technical novelty. Perhaps to satisfy reviewers, many papers

emphasize both complex models (addressed here) and fancy mathematics (see "Mathiness," in the next section of this article). While complex models are sometimes justified, empirical advances often come about in other ways: through clever problem formulations, scientific experiments, optimization heuristics, data-preprocessing techniques, extensive hyperparameter tuning, or by applying existing methods to interesting new tasks. Sometimes a number of proposed techniques together achieve a significant empirical result. In these cases, it serves the reader to elucidate which techniques are necessary to realize the reported gains.

Too frequently, authors propose many tweaks absent proper ablation studies, obscuring the source of empirical gains. Sometimes, just one of the changes is actually responsible for the improved results. This can give the false impression that the authors did more work (by proposing several improvements), when in fact they did not do enough (by not performing proper ablations). Moreover, this practice misleads readers to believe that all of the proposed changes are necessary.

In 2018 Gábor Melis, Chris Dyer, and Phil Blunsom demonstrated that a series of published improvements in language modeling, originally attributed to complex innovations in network architectures, were actually the result of better hyperparameter tuning.[27] On equal footing, vanilla LSTM (long short-term memory) networks, hardly modified since 1997, topped the leaderboard. The community may have benefited more by learning the details of the hyperparameter tuning

without the distractions. Similar evaluation issues have been observed for deep reinforcement learning[17] and generative adversarial networks.[24] See Sculley et al. for more discussion of lapses in empirical rigor and resulting consequences.[38]

In contrast, many papers perform good ablation analyses, and even retrospective attempts to isolate the source of gains can lead to new discoveries. Furthermore, ablation is neither necessary nor sufficient for understanding a method, and can even be impractical given computational constraints. Understanding can also come from robustness checks (as in Cotterell et al., which discovers that existing language models handle inflectional morphology poorly[9]), as well as qualitative error analysis.

Empirical study aimed at understanding can be illuminating even absent a new algorithm. For example, probing the behavior of neural networks led to identifying their susceptibility to adversarial perturbations.[44] Careful study also often reveals limitations of challenge data sets while yielding stronger baselines. A 2016 paper by Danqi Chen, Jason Bolton, and Christopher Manning studies a task designed for reading comprehension of news passages and finds that 73 percent of the questions can be answered by looking at a single sentence, while only 2 percent require looking at multiple sentences (the remaining 25 percent of examples were either ambiguous or contained coreference errors).[6] In addition, simpler neural networks and linear classifiers outperformed complicated neural architectures that had previously been evaluated on this task. In the same spirit, Rowan Zellers et

al. analyze and construct a strong baseline for the Visual Genome Scene Graphs data set in their 2018 paper.[45]

Mathiness

When writing a paper early in my Ph.D. program, I (Zachary Lipton) received feedback from an experienced post-doc that the paper needed more equations. The post-doc wasn't endorsing the system but rather communicating a sober view of how reviewing works. More equations, even when difficult to decipher, tend to convince reviewers of a paper's technical depth.

Mathematics is an essential tool for scientific communication, imparting precision and clarity when used correctly. Not all ideas and claims are amenable to precise mathematical description, however, and natural language is an equally indispensable tool for communicating, especially about intuitive or empirical claims.

When mathematical and natural-language statements are mixed without a clear accounting of their relationship, both the prose and the theory can suffer: problems in the theory can be concealed by vague definitions, while weak arguments in the prose can be bolstered by the appearance of technical depth. We refer to this tangling of formal and informal claims as *mathiness*, following economist Paul Romer, who described the pattern like this: "Like mathematical theory, mathiness uses a mixture of words and symbols, but instead of making tight links, it leaves ample room for slippage between statements in natural language versus formal language."[36]

Mathiness manifests in several ways. First, some papers abuse mathematics to convey technical depth—to bulldoze

rather than to clarify. Spurious theorems are common culprits, inserted into papers to lend authoritativeness to empirical results, even when the theorem's conclusions do not actually support the main claims of the paper. I (Steinhardt) was guilty of this in a 2015 paper with Percy Liang, where a discussion of "staged strong Doeblin chains" has limited relevance to the proposed learning algorithm but might confer a sense of theoretical depth to readers.[40]

The ubiquity of this issue is evidenced by the paper introducing the Adam optimizer.[19] In the course of introducing an optimizer with strong empirical performance, it also offers a theorem regarding convergence in the convex case, which is perhaps unnecessary in an applied paper focusing on non-convex optimization. The proof was later shown to be incorrect.[35]

A second mathiness issue is putting forth claims that are neither clearly formal nor clearly informal. For example, Yann Dauphin et al. argue that the difficulty in optimizing neural networks stems not from local minima but from saddle points.[11] As one piece of evidence, the work cites a statistical physics paper by Alan Bray and David Dean on Gaussian random fields and states that in high dimensions "all local minima [of Gaussian random fields] are likely to have an error very close to that of the global minimum."[5] (A similar statement appears in the related work of Anna Choromanska et al.[7]) This appears to be a formal claim, but absent a specific theorem it is difficult to verify the claimed result or to determine its precise content. Our understanding is that it is partially a numerical claim that the gap is small for typical settings

of the problem parameters, as opposed to a claim that the gap vanishes in high dimensions. A formal statement would help clarify this. Note that the broader interesting point in Dauphin et al. that minima tend to have lower loss than saddle points is more clearly stated and empirically tested.

Finally, some papers invoke theory in overly broad ways or make passing references to theorems with dubious pertinence. For example, the no-free-lunch theorem is commonly invoked as a justification for using heuristic methods without guarantees, even though the theorem does not formally preclude guaranteed learning procedures.

While the best remedy for mathiness is to avoid it, some papers go further with exemplary exposition. A 2013 paper by Léon Bottou et al. on counterfactual reasoning covers a large amount of mathematical ground in a down-to-earth manner, with numerous clear connections to applied empirical problems.[4] This tutorial, written in clear service to the reader, has helped to spur work in the burgeoning community studying counterfactual reasoning for ML.

Misuse of language

There are three common avenues of language misuse in machine learning: suggestive definitions, overloaded terminology, and suitcase words.

Suggestive Definitions

In the first avenue, a new technical term is coined that has a suggestive colloquial meaning, thus sneaking in connotations without the need to argue for them. This often manifests in anthropomorphic characterizations of

tasks (*reading comprehension* and *music composition)* and techniques (*curiosity* and *fear—I (Zachary) am responsible for the latter*). A number of papers name components of proposed models in a manner suggestive of human cognition (e.g., *thought vectors* and the *consciousness prior*). Our goal is not to rid the academic literature of all such language; when properly qualified, these connections might communicate a fruitful source of inspiration. When a suggestive term is assigned technical meaning, however, each subsequent paper has no choice but to confuse its readers, either by embracing the term or by replacing it.

Describing empirical results with loose claims of "human-level" performance can also portray a false sense of current capabilities. Take, for example, the "dermatologist-level classification of skin cancer" reported in a 2017 paper by Andre Esteva et al.[12] The comparison with dermatologists conceals the fact that classifiers and dermatologists perform fundamentally different tasks. Real dermatologists encounter a wide variety of circumstances and must perform their jobs despite unpredictable changes. The machine classifier, however, achieves low error only on IID (independent, identically distributed) test data.

In contrast, claims of human-level performance in work by Kaiming He et al.[16] are better qualified to refer to the ImageNet classification task (rather than object recognition more broadly). Even in this case, one careful paper (among many less careful) was insufficient to put the public discourse back on track. Popular articles continue to characterize modern image classifiers as "surpassing human abilities and effectively proving that

bigger data leads to better decisions," as explained by Dave Gershgorn,[13] despite demonstrations that these networks rely on spurious correlations, (e.g., misclassifying "Asians dressed in red" as ping-pong balls, reported by Pierre Stock and Moustapha Cisse[43]).

Deep-learning papers are not the sole offenders; misuse of language plagues many subfields of ML. Zachary Lipton, Alexandra Chouldechova, and Julian McAuley discuss how the recent literature on fairness in ML often overloads terminology borrowed from complex legal doctrine, such as *disparate impact*, to name simple equations expressing particular notions of statistical parity.[23] This has resulted in a literature where "fairness," "opportunity," and "discrimination" denote simple statistics of predictive models, confusing researchers who become oblivious to the difference and policymakers who become misinformed about the ease of incorporating ethical desiderata into ML.

Overloading Technical Terminology

A second avenue of language misuse consists of taking a term that holds precise technical meaning and using it in an imprecise or contradictory way. Consider the case of *deconvolution*, which formally describes the process of reversing a convolution, but is now used in the deep-learning literature to refer to *transpose convolutions* (also called *upconvolutions*) as commonly found in auto-encoders and generative adversarial networks. This term first took root in deep learning in a paper that does address deconvolution but was later overgeneralized to refer to any neural architecture using upconvolutions. Such overloading of terminology can create lasting confusion.

New ML papers referring to deconvolution might be (1) invoking its original meaning, (2) describing upconvolution, or (3) attempting to resolve the confusion, as in a paper by Caner Hazirbas, Laura Leal-Taixé, and Daniel Cremers,[15] which awkwardly refers to "upconvolution (deconvolution)."

As another example, *generative models* are traditionally models of either the input distribution p(x) or the joint distribution p(x,y). In contrast, discriminative models address the conditional distribution p(y|x) of the label given the inputs. In recent works, however, *generative model* imprecisely refers to any model that produces realistic-looking structured data. On the surface, this may seem consistent with the p(x) definition, but it obscures several shortcomings—for example, the inability of GANs (generative adversarial networks) or VAEs (variational autoencoders) to perform conditional inference (e.g., sampling from $p(x_2|x_1)$ where x_1 and x_2 are two distinct input features). Bending the term further, some discriminative models are now referred to as generative models on account of producing structured outputs, a mistake that I (Lipton), too, have made. Seeking to resolve the confusion and provide historical context, Shakir Mohamed and Balaji Lakshminarayanan distinguish between *prescribed* and *implicit* generative models.[30]

Revisiting batch normalization, Sergey Ioffe and Christian Szegedy describe *covariate shift* as a change in the distribution of model inputs.[18] In fact, *covariate shift* refers to a specific type of shift where, although the input distribution p(x) might change, the labeling function p(y|x) does not. Moreover, as a result of the influence of Ioffe and Szegedy, Google Scholar lists batch normalization as the

first reference on searches for "covariate shift."

Among the consequences of misusing language is the possibility (as with generative models) of concealing lack of progress by redefining an unsolved task to refer to something easier. This often combines with suggestive definitions via anthropomorphic naming. *Language understanding* and *reading comprehension*, once grand challenges of AI, now refer to making accurate predictions on specific data sets.

Suitcase Words

Finally, ML papers tend to overuse suitcase words. Coined by Marvin Minsky in the 2007 book *The Emotion Machine,*[29] suitcase words pack together a variety of meanings. Minsky describes mental processes such as consciousness, thinking, attention, emotion, and feeling that may not share "a single cause or origin." Many terms in ML fall into this category. For example, I (Lipton) note in a 2016 paper that *interpretability* holds no universally agreed-upon meaning and often references disjoint methods and desiderata.[22] As a consequence, even papers that appear to be in dialogue with each other may have different concepts in mind.

As another example, *generalization* has both a specific technical meaning (generalizing from train to test) and a more colloquial meaning that is closer to the notion of transfer (generalizing from one population to another) or of external validity (generalizing from an experimental setting to the real world). Conflating these notions leads to overestimating the capabilities of current systems.

Suggestive definitions and overloaded terminology can contribute to the creation of new suitcase words. In

the fairness literature, where legal, philosophical, and statistical language are often overloaded, terms such as *bias* become suitcase words that must be subsequently unpacked.

In common speech and as aspirational terms, suitcase words can serve a useful purpose. Sometimes a suitcase word might reflect an overarching aspiration that unites the various meanings. For example, *artificial intelligence* might be well suited as an aspirational name to organize an academic department. On the other hand, using suitcase words in technical arguments can lead to confusion. For example, in his 2017 book, *Superintelligence*, Nick Bostrom writes an equation (Box 4) involving the terms *intelligence* and *optimization power*, implicitly assuming that these suitcase words can be quantified with a one-dimensional scalar.[3]

SPECULATION ON CAUSES BEHIND THE TRENDS

Do the above patterns represent a trend, and if so, what are the underlying causes? We speculate that these patterns are on the rise and suspect several possible causal factors: complacency in the face of progress, the rapid expansion of the community, the consequent thinness of the reviewer pool, and misaligned incentives of scholarship vs. short-term measures of success.

Complacency in the face of progress

The apparent rapid progress in ML has at times engendered an attitude that *strong results excuse weak arguments*. Authors with strong results may feel licensed to insert arbitrary unsupported stories (see "Explanation

vs. Speculation" earlier in this article) regarding the factors driving the results; to omit experiments aimed at disentangling those factors (see "Failure to Identify the Sources of Empirical Gains"); to adopt exaggerated terminology (see "Misuse of Language"); or to take less care to avoid mathiness (see "Mathiness").

At the same time, the single-round nature of the reviewing process may cause reviewers to feel they have no choice but to accept papers with strong quantitative findings. Indeed, even if the paper is rejected, there is no guarantee the flaws will be fixed or even noticed in the next cycle, so reviewers may conclude that accepting a flawed paper is the best option.

Growing pains

Since around 2012, the ML community has expanded rapidly because of increased popularity stemming from the success of deep-learning methods. While the rapid expansion of the community can be seen as a positive development, it can also have side effects.

To protect junior authors, we have preferentially referenced our own papers and those of established researchers. And certainly, experienced researchers exhibit these patterns. Newer researchers, however, may be even more susceptible. For example, authors unaware of previous terminology are more likely to misuse or redefine language (see "Misuse of Language").

Rapid growth can also thin the reviewer pool in two ways: by increasing the ratio of submitted papers to reviewers and by decreasing the fraction of experienced reviewers. Less-experienced reviewers may be more likely

to demand architectural novelty, be fooled by spurious theorems, and let pass serious but subtle issues such as misuse of language, thus either incentivizing or enabling several of the trends described here. At the same time, experienced but overburdened reviewers may revert to a "checklist" mentality, rewarding more formulaic papers at the expense of more creative or intellectually ambitious work that might not fit a preconceived template. Moreover, overworked reviewers may not have enough time to fix—or even to notice—all of the issues in a submitted paper.

Misaligned incentives

Reviewers are not alone in providing poor incentives for authors. As ML research garners increased media attention and ML startups become commonplace, to some degree incentives are provided by the press ("What will they write about?") and by investors ("What will they invest in?"). The media provides incentives for some of these trends.

Anthropomorphic descriptions of ML algorithms provide fodder for popular coverage. Take, for example, a 2014 article by Cade Metz in *Wired*,[28] which characterizes an autoencoder as a "simulated brain." Hints of human-level performance tend to be sensationalized in newspaper coverage—for example, an article in the *New York Times* by John Markoff describes a deep-learning image-captioning system as "mimicking human levels of understanding."[25]

Investors, too, have shown a strong appetite for AI research, funding startups sometimes on the basis of a single paper. In my (Lipton) experience working with investors, they are sometimes attracted to startups

whose research has received media coverage, a dynamic that attaches financial incentives to media attention. Note that recent interest in chatbot startups co-occurred with anthropomorphic descriptions of dialogue systems and reinforcement learners both in papers and in the media, although it may be difficult to determine whether the lapses in scholarship caused the interest of investors or vice versa.

SUGGESTIONS

Suppose we are to intervene to counter these trends, then how? Besides merely suggesting that each author abstain from these patterns, what can we do as a community to raise the level of experimental practice, exposition, and theory? And how can we more readily distill the knowledge of the community and disabuse researchers and the wider public of misconceptions? What follows are a number of preliminary suggestions based on personal experiences and impressions.

For authors

We encourage authors to ask "What worked?" and "Why?" rather than just "How well?" Except in extraordinary cases, raw headline numbers provide limited value for scientific progress absent insight into what drives them. Insight does not necessarily mean theory. Three practices that are common in the strongest empirical papers are error analysis, ablation studies, and robustness checks (of, for example, choice of hyperparameters, as well as ideally of the choice of data set). These practices can be adopted by everyone, and we advocate their widespread use. For

some exemplar papers, consider the preceding discussion in "Failure to Identify the Sources of Empirical Gains." Pat Langley and Dennis Kibler also provide a more detailed survey of empirical best practices.[21]

Sound empirical inquiry need not be confined to tracing the sources of a particular algorithm's empirical gains; it can yield new insights even when no new algorithm is proposed. Notable examples of this include a demonstration that neural networks trained by stochastic gradient descent can fit randomly assigned labels.[46] This paper questions the ability of learning-theoretic notions of model complexity to explain why neural networks can generalize to unseen data. In another example, Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe explore the loss surfaces of deep networks, revealing that straight-line paths in parameter space between initialized and learned parameters typically have monotonically decreasing loss.[14]

When researchers are writing their papers, we recommend they ask the following question: *Would I rely on this explanation for making predictions or for getting a system to work?* This can be a good test of whether a theorem is being included to please reviewers or to convey actual insight. It also helps check whether concepts and explanations match the researcher's own internal mental model. On mathematical writing, we point the reader to Donald E. Knuth, Tracy Larrabee, and Paul M. Roberts's excellent guidebook, *Mathematical Writing*.[20]

Finally, being clear about which problems are open and which are solved not only presents a clearer picture to readers, but also encourages follow-up work and guards against researchers neglecting questions presumed

(falsely) to be resolved.

For Publishers and Reviewers

Reviewers can set better incentives by asking: "Might I have accepted this paper if the authors had done a worse job?" For example, a paper describing a simple idea that leads to improved performance, together with two negative results, should be judged more favorably than a paper that combines three ideas together (without ablation studies) yielding the same improvement.

Current literature moves fast at the expense of accepting flawed works for conference publication. One remedy could be to emphasize authoritative retrospective surveys that strip out exaggerated claims and extraneous material, change anthropomorphic names to sober alternatives, standardize notation, etc. While venues such as *Foundations and Trends in Machine Learning,* a journal from Now Publishers in Hanover, Massachusetts, already provide a track for such work, there are still not enough strong papers in this genre.

Additionally, we believe (noting our conflict of interest) that critical writing ought to have a voice at ML conferences. Typical ML conference papers choose an established problem (or propose a new one), demonstrate an algorithm and/or analysis, and report experimental results. While many questions can be approached in this way, when addressing the validity of the problems or the methods of inquiry themselves, neither algorithms nor experiments are sufficient (or appropriate). We would not be alone in embracing greater critical discourse: in NLP (natural-language processing), this year's COLING (Conference on Computational Linguistics) included a call

for position papers "to challenge conventional thinking."

There are many lines of further discussion worth pursuing regarding peer review. Are the problems described here mitigated or exacerbated by open review? How do reviewer point systems align with the values that we advocate? These topics warrant their own papers and have indeed been discussed at length elsewhere.

DISCUSSION

Folk wisdom might suggest not to intervene just as the field is heating up—you can't argue with success! We counter these objections with the following arguments: First, many aspects of the current culture are *consequences* of ML's recent success, not its *causes*. In fact, many of the papers leading to the current success of deep learning were careful empirical investigations characterizing principles for training deep networks. This includes the advantage of random over sequential hyperparameter search, the behavior of different activation functions, and an understanding of unsupervised pretraining.

Second, flawed scholarship already negatively impacts the research community and broader public discourse. The "Troubling Trends" section of this article gives examples of unsupported claims being cited thousands of times, lineages of purported improvements being overturned by simple baselines, data sets that appear to test high-level semantic reasoning but actually test low-level syntactic fluency, and terminology confusion that muddles the academic dialogue. This final issue also affects public discourse. For example, the European Parliament passed

a report considering regulations to apply if "robots become or are made self-aware."[10] While ML researchers are not responsible for all misrepresentations of our work, it seems likely that anthropomorphic language in authoritative peer-reviewed papers is at least partly to blame.

Greater rigor in exposition, science, and theory are essential for both scientific progress and fostering a productive discourse with the broader public. Moreover, as practitioners apply ML in critical domains such as health, law, and autonomous driving, a calibrated awareness of the abilities and limits of ML systems will help us to deploy ML responsibly.

Countervailing considerations

There are a number of countervailing considerations to the suggestions set forth in this article. Several readers of earlier drafts of this paper noted that *stochastic gradient descent tends to converge faster than gradient descent*—in other words, perhaps a faster, noisier process that ignores our guidelines for producing "cleaner" papers results in a faster pace of research. For example, the breakthrough paper on ImageNet classification proposes multiple techniques without ablation studies, several of which were subsequently determined to be unnecessary. At the time, however, the results were so significant and the experiments so computationally expensive to run that waiting for ablations to complete was perhaps not worth the cost to the community.

A related concern is that high standards might impede the publication of original ideas, which are more likely to be

unusual and speculative. In other fields, such as economics, high standards result in a publishing process that can take years for a single paper, with lengthy revision cycles consuming resources that could be deployed toward new work.

Finally, perhaps there is value in specialization: the researchers generating new conceptual ideas or building new systems need not be the same ones who carefully collate and distill knowledge.

These are valid considerations, and the standards we are putting forth here are at times exacting. In many cases, however, they are straightforward to implement, requiring only a few extra days of experiments and more careful writing. Moreover, they are being presented as strong heuristics rather than unbreakable rules—if an idea cannot be shared without violating these heuristics, the idea should be shared and the heuristics set aside.

We have almost always found attempts to adhere to these standards to be well worth the effort. In short, the research community has not achieved a Pareto optimal state on the growth-quality frontier.

Historical antecedents

The issues discussed here are unique neither to machine learning nor to this moment in time; they instead reflect issues that recur cyclically throughout academia. As far back as 1964, the physicist John R. Platt discussed related concerns in his paper on strong inference, where he identified adherence to specific empirical standards as responsible for the rapid progress of molecular biology and high-energy physics relative to other areas of science.[34]

There have been similar discussions in AI. As noted in the introduction to this article, Drew McDermott criticized a (mostly pre-ML) AI community in 1976 on a number of issues, including suggestive definitions and a failure to separate out speculation from technical claims.[26] In 1988, Paul Cohen and Adele Howe addressed an AI community that at that point "rarely publish[ed] performance evaluations" of their proposed algorithms and instead only described the systems.[8] They suggested establishing sensible metrics for quantifying progress, and analyzing the following: "Why does it work?", "Under what circumstances won't it work?", and "Have the design decisions been justified?"—questions that continue to resonate today.

Finally, in 2009 Timothy G. Armstrong and co-authors discussed the empirical rigor of information-retrieval research, noting a tendency of papers to compare against the same weak baselines, producing a long series of improvements that did not accumulate to meaningful gains.[1]

In other fields, an unchecked decline in scholarship has led to crisis. A landmark study in 2015 suggested that a significant portion of findings in the psychology literature may not be reproducible.[33] In a few historical cases, enthusiasm paired with undisciplined scholarship led entire communities down blind alleys. For example, following the discovery of X-rays, a related discipline on N-rays emerged before it was eventually debunked.[32]

CONCLUDING REMARKS

The reader might rightly suggest that these problems

are self-correcting. We agree. However, the community self-corrects precisely through recurring debate about what constitutes reasonable standards for scholarship. We hope that this paper contributes constructively to the discussion.

Acknowledgments

We thank the many researchers, colleagues, and friends who generously shared feedback on this draft, including Asya Bergal, Kyunghyun Cho, Moustapha Cisse, Daniel Dewey, Danny Hernandez, Charles Elkan, Ian Goodfellow, Moritz Hardt, Tatsunori Hashimoto, Sergey Ioffe, Sham Kakade, David Kale, Holden Karnofsky, Pang Wei Koh, Lisha Li, Percy Liang, Julian McAuley, Robert Nishihara, Noah Smith, Balakrishnan "Murali" Narayanaswamy, Ali Rahimi, Christopher Ré, and Byron Wallace. We also thank the ICML Debates organizers for the opportunity to work on this draft and for their patience throughout our revision process.

References

1. Armstrong, T. G., Moffat, A., Webber, W., Zobel, J. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* 601-610.
2. Bengio, Y. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, ed. G. Montavon, G. B. Orr, KR Müller, 437-78. *Lecture Notes in Computer Science* 7700. Springer, Berlin, Heidelberg.

3. Bostrom, N. 2017. *Superintelligence*. Paris: Dunod.
4. Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D.X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., Snelson, E. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *The Journal of Machine Learning Research* 14(1), 3207–3260.
5. Bray, A. J., Dean, D. S. 2007. Statistics of critical points of Gaussian fields on large-dimensional spaces. *Physical Review Letters* 98(15), 150201; https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.98.150201.
6. Chen, D., Bolton, J., Manning, C. D. 2016. A thorough examination of the CNN/*Daily Mail* reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2358–2367.
7. Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., LeCun, Y. 2015. The loss surfaces of multilayer networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics.*
8. Cohen, P. R., Howe, A. E. 1988. How evaluation guides AI research: the message still counts more than the medium. *AI Magazine* 9(4), 35.
9. Cotterell, R., Mielke, S. J., Eisner, J., Roark, B. 2018. Are all languages equally hard to language-model? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2.
10. Council of the European Union. 2016. Motion for a European Parliament Resolution with Recommendations to the Commission on Civil Law Rules on Robotics; http://www.

europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.
11. Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, 2933–2941.
12. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639), 115-118.
13. Gershgorn, D. 2017. The data that transformed AI research—and possibly the world. Quartz; https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.
14. Goodfellow, I. J., Vinyals, O., Saxe, A. M. 2015. Qualitatively characterizing neural network optimization problems. In *Proceedings of the International Conference on Learning Representations*.
15. Hazirbas, C., Leal-Taixé, L. Cremers, D. 2017. Deep depth from focus arXiv: 1704.01085; https://arxiv.org/abs/1704.01085.
16. He, K., Zhang, X., Ren, S., Sun, J. 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026-1034.
17. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D. 2018. Deep reinforcement learning that matters. In *Proceedings of the 32nd Association for the*

Advancement of Artificial Intelligence Conference.
18. Ioffe, S. Szegedy, C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* 37; http://proceedings.mlr.press/v37/ioffe15.pdf.
19. Kingma, D. P., Ba, J. 2015. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.
20. Knuth, D. E., Larrabee, T., Roberts, P. M. 1987. Mathematical writing; http://jmlr.csail.mit.edu/reviewing-papers/knuth_mathematical_writing.pdf.
21. Langley, P., Kibler, D. 1991. The experimental study of machine learning; http://www.isle.org/~langley/papers/mlexp.ps.
22. Lipton, Z. C. 2016. The mythos of model interpretability. International Conference on Machine Learning Workshop on Human Interpretability.
23. Lipton, Z. C., Chouldechova, A., McAuley, J. 2017. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, 8136-8146. arXiv: 1711.07076; https://arxiv.org.abs/1711.07076.
24. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O. 2017. Are GANs created equal? A large-scale study. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*. arXiv:1711.10337. https://arxiv.org.abs/ arXiv:1711.10337.
25. Markoff, J. 2014. Researchers announce advance in image-recognition software. *New York Times* (November 17); https://www.nytimes.com/2014/11/18/

science/researchers-announce-breakthrough-in-content-recognition-software.html.
26. McDermott, D. 1976. Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin* 57, 4–9.
27. Melis, G., Dyer, C., Blunsom, P. 2018. On the state of the art of evaluation in neural language models. In *Proceedings of the International Conference on Learning Representations*.
28. Metz, C. 2014. You don't have to be Google to build an artificial brain. *Wired* (September 26); https://www.wired.com/2014/09/google-artificial-brain/.
29. Minsky, M. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster.
30. Mohamed, S., Lakshminarayanan, B. 2016. Learning in implicit generative models. arXiv:1610.03483; https://arxiv.org.abs/ 1610.03483
31. Noh, H., Hong, S., Han, B. 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the International Conference on Computer Vision*, 1520-1528.
32. Nye, M. J. 1980. N-rays: an episode in the history and psychology of science. *Historical Studies in the Physical Sciences* 11(1), 125–56.
33. Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251): aac4716.
34. Platt, J. R. 1964. Strong inference. *Science* 146 (3642), 347-353.
35. Reddi, S. J., Kale, S., Kumar, S. 2018. On the convergence of Adam and beyond. In *Proceedings of the International*

Conference on Learning Representations.

36. Romer, P. M. 2015. Mathiness in the theory of economic growth. *American Economic Review* 105(5), 89–93.
37. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A. 2018. How does batch normalization help optimization? (No, it is not about internal covariate shift). In *Proceedings of the 32nd Conference on Neural Information Processing Systems*; https://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.pdf.
38. Sculley, D., Snoek, J., Wiltschko, A., Rahimi, A. 2018. Winner's curse? On pace, progress, and empirical rigor. In *Proceedings of the 6th International Conference on Learning Representations, Workshop Track*.
39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1), 1929–1958; https://dl.acm.org/citation.cfm?id=2670313.
40. Steinhardt, J., Liang, P. 2015. Learning fast-mixing models for structured prediction. In *Proceedings of the 32nd International Conference on Machine Learning* 37, 1063–1072; http://proceedings.mlr.press/v37/steinhardtb15.html.
41. Steinhardt, J., Liang, P. 2015. Reified context models. In *Proceedings of the 32nd International Conference on Machine Learning* 37, 1043-1052; https://dl.acm.org/citation.cfm?id=3045230.
42. Steinhardt, J., Koh, P. W., Liang, P. S. 2017. Certified defenses for data poisoning attacks. In *Proceedings of the 31st Conference on Neural Information Processing Systems*; https://papers.nips.cc/paper/6943-certified-

defenses-for-data-poisoning-attacks.pdf.
43. Stock, P., Cisse, M. 2017. ConvNets and ImageNet beyond accuracy: explanations, bias detection, adversarial examples and model criticism. arXiv:1711.11443; https://arxiv.org/abs/1711.11443.
44. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. 2013. Intriguing properties of neural networks. International Conference on Learning Representations. arXiv:1312.6199; https://arxiv.org/abs/ 1312.6199.
45. Zellers, R., Yatskar, M., Thomson, S. Choi, Y. 2018. Neural motifs: scene graph parsing with global context. In *Computer Vision and Pattern Recognition,* 5831-5840.
46. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations*.

Zachary Lipton *is an Assistant Professor at Carnegie Mellon University in the Tepper School of Business with appointments in the Machine Learning Department and the Heinz School of Public Policy. He also collaborates closely with Amazon, where concurrent with his final year of Ph.D. work, he helped to grow AWS' Amazon AI team into a large applied research organization and contributed to the Apache MXNet deep learning framework. His work addresses core technical challenges and real-world applications of machine learning (ML), focusing on the robustness of ML systems, applications to healthcare, and the real-world behavior and social impacts of deployed algorithms. He is the founding editor of*

the Approximately Correct blog and an author of Dive into Deep Learning, an interactive open-source book teaching deep learning through Jupyter notebooks. Find him on the web (zacklipton.com), Twitter (@zacharylipton) or GitHub (@zackchase).

Jacob Steinhardt *recently finished a Ph.D. at Stanford University and will be joining UC Berkeley as an assistant professor of statistics. His research focuses on making the conceptual advances necessary for machine learning systems to be reliable and aligned with human values. He has also collaborated with policy researchers to understand and avoid potential misuses of machine learning. He is also a technical advisor for the Open Philanthropy Project.*