Pitfalls in Machine Learning Research: Reexamining the Development Cycle

Stella Biderman
The AI Village
Booz Allen Hamilton
stellabiderman@gmail.com

Walter J. Scheirer
The AI Village
University of Notre Dame
walter.scheirer@nd.edu

Abstract

Applied machine learning research has the potential to fuel further advances in data science, but it is greatly hindered by an *ad hoc* design process, poor data hygiene, and a lack of statistical rigor in model evaluation. Recently, these issues have begun to attract more attention as they have caused public and embarrassing issues in research and development. Drawing from our experience as machine learning researchers, we follow the applied machine learning process from algorithm design to data collection to model evaluation, drawing attention to common pitfalls and providing practical recommendations for improvements. At each step, case studies are introduced to highlight how these pitfalls occur in practice, and where things could be improved.

1 Introduction

There is much to be excited about in the field of machine learning these days. From championship video game AI to advances in autonomous vehicles, it seems that no matter where we look we see machine learning transforming yet another domain. But while the field has been taking its victory lap for these successes, problems have surfaced from the depths of the machine learning development cycle that potentially jeopardize the entire endeavor by producing illusory experimental effects. In some cases these problems are systematic, and extend beyond the existing discussion of problems related to dataset bias. Without an adequate response to solve them, the utility of machine learning as a constructive technology is brought into question.

Pitfalls can emerge at three critical points in the applied machine learning development cycle:

1. In the design process of an algorithm, when a team is formed to solve a problem, assumptions about the problem and solution are formulated, and an algorithm is developed.
2. At the point of data collection for a model that will be trained using the newly developed algorithm.
3. During evaluation, where the model's performance as a solution to the problem is assessed.

There are serious technical and ethical ramifications when there is a breakdown at any critical point in the development cycle. Illusory experimental effects lead to machine learning papers that not only fail to replicate, but also create serious and often difficult-to-diagnose problems for the people who try to implement the research in practice. Recent examples of this include algorithms for determining criminal tendencies [53, 22] and detecting sexual orientation [51] from photos of faces, as well as commercial services that automatically determine suitable job candidates for hiring managers [12]. In all of these cases, the claimed functionality is not what it appears to be due to underlying problems in experimental design, data use, and evaluation [14, 3, 4].

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

It is important to note that we are not merely claiming that some research is morally dubious. While some of the research we discuss in this paper *is* morally dubious, we see the same problems in research that isn't, including neural architecture search, the ImageNet dataset, and route-finding algorithms. Problems in the machine learning development cycle also appear when machine learning is applied to other fields of science. For example, Liu et al. [30] found that out of tens of thousands of papers on deep learning published in the medical imaging literature, only a tiny fraction were methodologically sound.

The objective of this paper is to identify and illustrate common pitfalls in the machine learning development cycle. It is not to shame or embarrass particular researchers or organizations. For every example we cite in this paper, there are a dozen other examples that could have been given. Furthermore, all of the pitfalls raised in this paper are problems that have come up in our own work. The first author has withdrawn submitted papers after issues we raise were brought to her attention and has had to explain to end-users that models they developed could not be put into practice due to methodological shortcomings. The second author has worked on data-driven modeling for years, and regrets not always scrutinizing data sources and delivering evaluations that could have been more rigorous.

While some of the issues are things that individual researchers can address in their work, many are not. Even when the problems can be addressed individually, the culture of machine learning research fails to enforce important behavioral norms. We aim to fix that.

The bulk of the rest of this paper traces the development cycle of machine learning research. We begin by discussing algorithm design, before moving on to data collection and finally model evaluation. Within each topic we raise some of the pressing problems with current machine learning practice, elucidate these problems with case studies, and then present recommendations to improve the current practice of the field. In the final section of this paper we discuss takeaways, with a focus on things the reader can

2 Designing the Right Algorithm

Before one can answer the question "Am I doing the research right?" one must first tackle the question "Am I doing the right research?" Although ethical issues with machine learning research have attracted increasing attention recently, we contend that many projects suffer from fundamental design flaws that make them — even on a purely technical level — non-starters from the beginning.

2.1 Problems with Algorithm Design

Not Engaging with Stakeholders. Modern machine learning and data science has national and even global repercussions, but researchers rarely admit the scope of the interests involved. While progress has been made in getting researchers and engineers to identify users as stakeholders, as machine learning and data science technologies become increasingly popular for public use the importance of recognizing anyone who may interact with the system as a stakeholder becomes crucial. For example, the stakeholders in FBI predictive policing algorithms [19] include every person who will set foot in the U.S. while the algorithms are in use and the stakeholders in the routing algorithms in Google Maps and Waze include the people whose neighborhoods and lives are disrupted by having drivers rerouted by their homes [31, 29]. Existing notions of participatory design in machine learning are not sufficient to address these sorts of situations.

Ignoring Assumptions. Algorithms necessarily rely on oversimplified models of the world, but not enough attention is paid to the assumptions that are designed into models and the impacts that they have on the results. This is a perpetual problem in computer science, famously highlighted a decade ago in the blog post "Falsehoods Programmers Believe About Names" [32]. Assumptions that are built into algorithms and datasets (often implicitly) can influence the results of research in ways that undermine its credibility [24, 48] or dissuade people from pursuing certain types of research [5, 10].

Lack of Oversight. When most scientists seek to do a study that impacts humans, they are required to obtain approval by an Institutional Review Board (IRB) whose job it is to protect the rights and interests of the subjects of the study. Unfortunately, most applied machine learning and data science research is exempt from requiring IRB approval because it analyzes preexisting data. This often means

that machine learning research is not reviewed by a third party until it is given to peer-reviewers, a point far too late to make meaningful changes to the design of the experiment or algorithm.

2.2 Case study: Gender and Machine Learning

Automated gender recognition (AGR) is a textbook example of how assumptions made in the design process can invalidate research. One example of this is how they operationalize gender without attending to the direction of causation. There are hundreds of papers that claim to be able to accurately determine people's gender based on anything from a writing sample [6, 56, 20], to their gait [57, 16, 18], to the texture of their skin [54, 27, 1]. Regardless of what one thinks constitutes "gender," almost nobody believes that one's gender is determined by their handwriting, gait, or skin care regimen. Indeed, many of the study authors explicitly acknowledge this disconnect before proceeding to ignore it.

The typical justifications of these methodologies uses a V-shaped causation pattern: something causes gender, that same thing causes the attribute measured, and therefore we can infer things about gender from the measurement. This is invalid reasoning, but it is highly prevalent in machine learning research. When these algorithms are being used to prescriptively assign gender labels to people, something that all of the mentioned papers cite as an application of their work, it is not sufficient to invoke correlation instead of causation as a justification of the effect. None of the referenced papers employ sufficient controls to make any sort of causal claims, but every paper makes them.

AGR research also highlights how ontological assumptions can interfere with research. According to Keyes [24], 94.8% of AGR papers define gender in a binary fashion with no acknowledgment of the existence of transgender people. By prescriptively defining the bins that people must fall into, the algorithm is incapable of understanding humans beyond the scope of the author's conceptions. This is a very common way for researchers to imprint their own biases into their data, and best practices in fields such as medicine [7] and HCI [45] strongly discourage it.

2.3 Best Practices: Recommendations for Algorithm Design

Not every problem needs to be solved via an algorithm. Although as machine learning researchers we are inherently biased towards trying to solve problems with machine learning, many issues can be solved by simply *not trying to solve them at all*. In a conversation about whether or not Twitter's automatic cropping algorithm is racially biased, Amy Zhang [59] pointed out

> The easiest fix for that biased cropping AI? No it's not to build another AI – it's to give people the power to select crop boundaries when posting a photo.

A similar sentiment shared by many others. Twitter took this criticism to heart, and announced changes to cropping including giving users control over how their images were cropped [2]

Democratize your notions of stakeholdership. Any person who produces data used in a model, uses a model, or whose life is impacted by the outputs of a model is a stakeholder. This especially includes groups who are[1] undervalued and underrepresented in the design process.

Integrate stakeholder feedback into the default ML development cycle. While the recent paper "Participation is not a Design Fix for Machine Learning" [44] raises important points about the limits of how participatory design is often done, that does not mean that participatory design is not a crucial component of the ML development cycle. It is vital that machine learning researchers take the criticisms of their research by stakeholders seriously, making meaningful changes to their concepts and approaches rather than "participation-washing" their research. Bhatt et al. [9] and Sloane et al. [44] provide actionable recommendations for researchers looking to improve their next study.

Subject research to ethical review. Recent strides towards improving the ethicality of machine learning research have been made by NeurIPS's requirement of a Broader Impacts statement. While this is a good first step, ethical review needs to be taken more seriously industry-wide, and integrated earlier in the research process to prevent significant amounts of time and money being spent on fundamentally invalid or harmful research. Even when the problems with research later comes out (e.g., predicting

[1] It is typical to say "who are historically" here, but this is misleading and casts diversification as a solved problem which it is not.

criminality from facial structure, training language models with hate speech) that doesn't prevent the harm done when people take this research and base deployed AI algorithms on it.

3 Collecting the Data

The adage "what is counted counts" has never been more true than it is today. Troves of data are collected and stored every day, much of it to be analyzed by machine learning algorithms. How that data is collected and what precisely it counts is vital to understanding the resulting analysis. Yet far too often questions of data methodology are ignored by industry and academic researchers alike.

3.1 Problems with Data Collection

Statistical and Social Bias. At this point, it should not be surprising to anyone working on machine learning to hear that datasets are heavily biased. However, bias remains a key problem facing all fields pursuing data-driven modeling. Biases tend to manifest in one of two ways: statistical and social. Statistical biases can take the form of class imbalance or repeating irregularities associated with both salient and non-salient information in data points. While recent discussion makes this seem as though it were a newly discovered problem, statistical bias in machine learning has been studied for decades [26]. Social biases can be more subtle and tend to occur as artifacts of a data collection process or as reflections of societal biases.

Not Testing on Data Collected from the Real World. The machine learning community has decided that widely-used benchmark datasets are the best way to evaluate an algorithm's performance. Unfortunately this can cause assessments of an algorithm's reliability in a paper to be wildly different from what happens in the real world [17]. The only way to evaluate an algorithm's performance in the real world is on freshly collected real world data from the population that the algorithm will actually be applied to. Machine learning can look to the field of robotics for best practices, as it has always done this and never hesitates to point out the mismatch between datasets and the real world [46].

The Limitations of the Big Data Paradigm. The dominant strategy in AI product development has been to collect vast troves of data for high-capacity machine learning models. There can be no doubt that big data has led to advances in numerous applications, from photo tagging to autonomous vehicles. But this approach has its limitations, some of which are only now becoming apparent. Perhaps obvious in retrospect, "bigness" can be a liability. For many datasets, there are too many images for manual scrutiny or meaningful validation. Researchers have no good way to exhaustively examine each and every datum, which can lead to problematic data samples being provided to an algorithm during training. Recent work by Prabhu and Birhane [39] highlights major issues with several commonly used image datasets, and how to address them in research.

Noisy Labeling. Because we lean heavily on human annotators for supervised learning, what is inherently a noisy labeling process is prone to mistakes. Rater reliability can be assessed [41, 35], but rarely is. Nearly all datasets in common use have a single label assigned by one person to each sample. This does not give us any indication about the correctness or difficulty of the sample, and can lead to problems down the line during training. The labeling process is also prone to malicious attack. It has been shown that it is possible to change labels in such a way that provides a favorable outcome to an attacker (e.g., a backdoor in a trained model) [13]. The problem is exacerbated by big data: if a small number of labels change in a sea of millions of samples, does anybody notice? Those changes may be consequential to model outcomes.

3.2 Case Study: The Tiny Images Dataset

A very recent example that reflects almost all of the above problems is the retraction [48] of the Tiny Images dataset [50], which had been used for object recognition research. This dataset contains very small images (32×32 pixels) for over 50,000 different noun categories, and was meant to develop visual recognition capabilities that match the "remarkable tolerance of the human visual system." As of this writing, the paper describing the dataset has been cited over 1,700 times, and numerous algorithms have been developed using it as source data. The retraction was prompted by an investigation by Prabhu and Birhane [39], who noted that Tiny Images used several categories for images labeled with racial and misogynistic slurs. Under closer scrutiny, they found that the dataset

also contained non-consensual pornography such as up-skirt photographs and imagery degrading to various marginalized groups.

The problems that were exposed in Tiny Images map directly to the problems we have singled out related to bias, big data, and labeling. Tiny Images most obviously suffered from a case of social bias in its racist and misogynistic categories. These were unfortunate reflections of Internet culture, which were unavoidable in the collection strategy used by the creators of the dataset: an automated data procedure that relied on nouns from WordNet [33]. This saga also exemplifies the limitations of the big data paradigm. According to the retraction issued by Torralba et al. "The dataset is too large (80 million images) and the images are so small (32 x 32 pixels) that it can be difficult for people to visually recognize its content." Thus, it is argued that the very advantage of big data is rendered moot by the presence of even a small number of problematic data samples. Because it is not possible to find all of the problematic instances in a dataset, according to Torralba et al., the only recourse is to take a dataset out of service if problems are found. And when it comes to labels, Tiny Images contained arbitrary label assignments that reflected accepted and derogatory racial categories in WordNet. Moreover, the crawling process relied on tags from the web, not multiple assignments from a collection of annotators. Without a consensus judgment, this means the accuracy of the labels remains unclear for many of the images. Finally, there is a temporal aspect to the labels: assigned labels can evolve over time as social and cultural norms change.

3.3 Best Practices: Recommendations for Data Collections

Hypothesis Driven Data Collection. Instead of simply hoovering up data and then trying to use it for whatever arbitrary application that comes down the line, a better approach would be to design a collection with a hypothesis in mind. There is no reasonable expectation that any question can be answered by using a generic pool of data, no matter how large. Across the natural sciences, experimental data collection is tightly coupled with a specific hypothesis that is formulated before work begins. The same should be true of experiments in machine learning.

Auditing and Documenting Datasets. While we acknowledge that it is impossible to go exhaustively through today's machine learning datasets by hand, there are still some sensible strategies for auditing that can be used. Given that most dataset come from sites where users can upload their own content, there are fairly obvious problems to look for in a targeted way: hate speech, profanity, and pornography. Less obvious catches can be made by auditing the sources of the data. Are celebrity news sites overly represented in a dataset for photo captioning? That may lead to racial bias in operation [42]. Thus better heterogeneity in sourcing is needed. In software engineering, a set of tests that is not comprehensive, but still useful to reveal failure modes is known as *smoke testing*. This idea transfers nicely to dataset auditing, where feasible checks for the above items and others can made in a reasonable amount of time. From the point of view of documentation, the "Datasheets for Datasets" [21] framework is something that people are beginning to use and which would benefit the field if adopted as a standard [43, 15, 47].

Quantify Annotator Uncertainty. When it comes to the quantification of annotator uncertainty, the recommendation here is to quantify aleatoric uncertainty [25, 23]. This is the uncertainty estimated and attempted to be removed when aggregating data from different annotators. With a quantitative value of uncertainty for a datapoint, a decision can be made to use or not use that point, or perhaps weight its influence appropriately, which has been shown to be effective [37]. The process can include an assessment of how familiar annotators are with the domain, as well the social and cultural norms associated with the type of data being annotated.

Dataset Revision Process. The above problems mean that at some point in a dataset's life-cycle, it will need to revised. A revision of a dataset can remove problematic information, document what has been removed from the previous version, and provide an explanation for why the revision was necessary. This may not completely remove all problems from a dataset, especially in a big data context, but it is a way to address specific problems as they are raised. Revision control systems for data should be developed to ease this process. Importantly, a standard for dataset revision should be defined and adopted by the community. "Datasheets for Datasets" [21] is a framework that has attracted attention for this, but researchers and practitioners have been slow to put it into practice.

4 Evaluating the Model

The field of machine learning is founded upon dataset-based evaluation. On the one hand, using standard datasets provide a common basis for comparison across different algorithms, as well as sufficient data for self-contained evaluations. On the other hand, because datasets are self-contained worlds they often misrepresent algorithm performance and can result in misleading findings if datasets are too heavily relied on. Exacerbating these problems are intentional or unintentional misuses of learning algorithms. Here we make specific recommendations on how evaluation can be improved informed by common practices in other experimental fields and previous observations from the machine learning literature that have gone unheeded. These recommendations are made for artificial neural networks, but in some cases can apply to other learning algorithms as well.

4.1 Problems with Model Evaluation

Lack of Rigorous Statistical Evaluation. Statistically verifying the results of a scientific paper is essential to its publication and the credibility of its results. Despite the attention to rigor in the algorithm development process, many of the most popular and successful machine learning models do not apply statistically rigorous techniques in evaluation. This can take many forms, including failing to report interval estimates for results, failing to analyze variance and random seed effects, and failing to properly control for covariates such as training methodology.

A machine learning model is usually considered successful if it can exceed state-of-the-art performance on standard datasets, but researchers rarely pay attention to the statistical significance of their results. Making judgments based on a small number of data points with little attention to the significance of their findings leads researchers to incorrectly over- or under-value work. In the worst cases, the lack of significance testing leads to a failure to notice that methodologies do not beat basic null models for the task.

A largely unacknowledged problem in neural network experiments is the lack of k-fold testing by varying the value of the random seed during training. Rarely do we find a paper that reports error based on this form of evaluation. It is well known that different random initializations of the elementary parameters can lead to drastically different results [36]. For datasets with fixed training, validation, and testing partitions, this presents a dilemma. An honest experimenter may get lucky or unlucky, depending on the choice of the seed. Thus the reported results may not reflect what happens on average for a series of training runs. A dishonest experimenter may attempt to mine seeds for an extended period of time, looking for a favorable starting place that leads to good results on the test set.

Failing to Compare to Null Models. Not all accuracies are created equal. On some tasks, e.g., malware detection, getting to 90% accuracy is trivial while on others, e.g., predicting the outbreak of war, it would be world-changing. While this fact is well-known to researchers, it is not sufficiently respected by them.

A recent example of this is "Criminality from Face" research, where deep learning was reported to be able to determine whether or not somebody is a criminal based on a photo of their face [53, 22]. Bowyer et al. [11] demonstrated that accurate results can be achieved for this task because of the organization of the datasets used: mugshot photos from government data sources are labeled "criminal" and ordinary public photos crawled from the web are labeled "non-criminal." The authors found that the results were completely explained by dataset classification. In the now classic paper an "Unbiased Look at Dataset Bias," Torralba and Efros [49] warned about this very problem. That paper has largely been remembered for kicking off the study of individual biases within computer vision datasets, but it made a more important observation about datasets as self-contained worlds: they possess a certain visual style at a global-level, which can easily be learned. If an experimenter is labeling the datasets to suit their needs, either intentionally or unintentionally, this observation can be exploited.

Lack of 3rd Party Evaluation. The academic peer review process is meant to verify the scientific integrity of work, which includes reproducibility. Earlier in this paper, we noted that the vast majority of published deep learning models for medical image analysis cannot be independently verified [30]. It is now standard practice by corporate research labs to not publish code or data with their papers in the interest of protecting intellectual property. This means that it is impossible to verify the claims

made without re-implementing the work described in the paper from scratch, which is sometimes impossible if corresponding configurations are not made available to the public, or if details have been omitted. Even when code is available, replications of machine learning research often fail [40].

Aggravating this problem has been the dramatic increase in papers submitted to machine learning-oriented publication venues. CVPR alone went from 2,123 papers in 2015 to 6,424 in 2020. Reviewers, who are overburdened even with the average load of 5-6 papers, do not have time to work with any available code under these conditions.

On the commercial side, companies are under no obligation to submit their products to peer review. In other industries, it is common practice to seek 3rd party evaluation of a product for quality or standards certification. This has rarely been the case in the AI industry. When blackbox machine learning products are released without 3rd party evaluation, problems that might otherwise have been identified in testing can emerge in operation. In one example, Microsoft released a blackbox web app for photo captioning that reproduced racial stereotypes which are prevalent in computer vision datasets composed of celebrity photos [42]. In another case, Twitter's blackbox photo cropping algorithm demonstrated racial bias, and led someone to start a public experiment on the platform [38]. Presently, we only see change when there is large public outcry, which isn't a sustainable strategy.

4.2 Case Study: Neural Architecture Search and Randomly Wired Neural Networks

Neural architecture search (NAS) is a field of deep learning that tries to optimize the neural network architecture and find networks that produce better results when trained. Unfortunately neural architecture search by and large does not work, and many NAS researchers seem to have not noticed this fact due to poor statistical practice.

The landmark paper by Xie et al. [55] points out that that NAS papers typically compare against each other with no external reference points or null models to compare to. To address this gap, they train neural networks whose computational graphs are generated randomly. To decrease the influence of any bias or knowledge that the authors have on the null models, they decide to generate graphs using standard random graph generators from social network analysis. These random graph generators have the additional benefit of having never been applied to NAS before. The surprising result was that some of the random network generators achieved "state-of-the-art" performance.

Unfortunately, the neural architecture search community does not seem to have learned much from — or even understood — Xie et al. [55]. At the time of writing Xie et al. [55] had 144 citations. Of those, more than 75% of the papers cite it as an example of neural architecture search being effective, with more than 25% *explicitly describing it as a state-of-the-art methodology*. Despite the fact that hundreds of NAS papers have been published since Xie et al. [55], subsequent research has continued to validate the fact that basic null models based on random search remain the "state-of-the-art" for NAS [28, 52, 34]. This term itself is a bit of a misnomer though, as any methodology that fails to outperform random networks is at a very basic level failing to do optimization at all.

Ottelander et al. [34] look into the problems with NAS research deeper, replicating eight recent NAS systems on five image datasets. They find evidence that all of the problems identified in section 4.1 are widespread in NAS:

> **Lack of Statistical Rigor:** Poor statistical rigor has allowed for confounding factors to cast doubt on or invalidate results on widely used search spaces. In particular, they find that the Differentiable Architecture Search (DARTS) space is not capable of producing real NAS innovation as the exact training set-up and choice of hyperparameters cause a significantly large variation in results than architecture improvements.
>
> **Lack of Null Models:** None of the papers in question rigorously compare to adequate null models, and all NAS methods examined either do not improve over null models or do not significantly do so.
>
> **Lack of Replicability:** While the eight NAS systems they replicated were selected because they had open source code, the authors note that this is not common. Additionally, they discuss that many NAS papers are non-comparable due to using very different search spaces.

4.3 Best Practices for Model Evaluation

Statistically Validate Results. A better option than reporting results with just a point estimate is to report the average and standard error of the estimator. This can be done by varying the random seed, subsampling the test dataset, or by hypothesis testing. While some of these techniques are catching on, cultural norms around meaningful levels of statistic rigor in published research are needed.

Train a Dataset Classifier as a Control Model. In order to determine whether or not a dataset or data source is being learned instead of the intended function, our recommendation is to perform a control experiment that swaps the target labels in the training set for dataset labels. This will lead to the creation of a dataset classifier at training time if a dataset is being learned. This process is especially important in cases where multiple datasets are being used in the training set. If this detector is able to correctly identify the source of new instances, it is likely learning the statistical differences between datasets rather than the property of interest.

Conduct Third Party Evaluation. Authors of machine learning papers deserve another set of eyes on their work. Conferences and journals should give reviewers access to the exact code, data and configuration for the experiments described in papers introducing new algorithms. Area chairs and area editors should expect reviewers to run and tinker with the code, and referee reports should include an analysis of reviewer experiments focusing on replication and sensitivity to free parameters (including a change of data). This will inevitably slow down the rate of publication, but improve the quality of published work (and perhaps ease the burden on the conference reviewing process). It is very much in-line with recent calls for machine learning to join the Slow Science movement [8]. With respect to commercial products, an organization similar to Underwriters Laboratories (`https://www.ul.com/`) should be established to certify the safety and correctness of machine learning products. There is already some precedent for this in computer security [58].

5 Takeaways for Doing Better Data Science

The central challenge for methodologists is to convince practitioners that their suggestions for improvement are worth pursuing. Resistance to changing methodologies that are perceived as working is quite reasonable – as one reviewer of this paper put it, "Hypothesis-driven data collection: this sounds good on paper, but what does it mean in practice... Even if cost were no issue, ML has made enormous progress using standard benchmark datasets, such as ImageNet, to allow for objective comparison of results from different labs using different methods. Are the authors suggesting that the field abandon this approach?"

While we agree that the methodologies commonplace in applied machine learning research has produced great success, that doesn't mean that we should not continue to strive to do better. To this point, we have sought to not only criticize current scientific practices but also to elevate work that exemplifies or instructs research to a higher standard. We hope that this, together with our analysis of prominent failure cases at each stage in the development process, inspires researchers to hold themselves and each other to higher standards.

If there is one thing the reader takes away from this paper, we hope it is that there are concrete steps that individuals bring home with them to improve their methodological practices. This includes thinking about whether or not a problem needs to be solved via an algorithm, applying more rigorous standards of statistical analysis, auditing and documenting datasets, and comparing results to null models. These recommendations can be implemented on a project-by-project basis without significant external support. Additionally, although many of our suggestions require larger scale changes than one person can accomplish alone, you can encourage the development of the necessary norms and cultural attitudes by asking these questions in peer review, posing them to project leads, and making your personal commitment to them known. Even if people do not yield to your questions, raising them and normalizing them as a point of discussion is an important first step.

Acknowledgments

We would like to thank the AI Village for invaluable discussion of the themes of this paper and feedback on the manuscript.

References

[1] Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 78(15):20835–20854, 2019.

[2] Parag Agrawal and Dantley Davis. Transparency around image cropping and changes to come. *Twitter Blog*, 2020.

[3] Blaise Agüera y Arcas, Alexander Todorov, and Margaret Mitchell. Do algorithms reveal sexual orientation or just expose our stereotypes? *medium.com*, 2018.

[4] Ifeoma Ajunwa. Automated employment discrimination. *Available at SSRN 3437631*, 2019.

[5] Kendra Albert, Jon Penney, Bruce Schneier, and Ram Shankar Siva Kumar. Politics of adversarial machine learning. In *Towards Trustworthy ML: Rethinking Security and Privacy for ML Workshop, Eighth International Conference on Learning Representations (ICLR)*, 2020.

[6] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[7] Greta R Bauer, Jessica Braimoh, Ayden I Scheim, and Christoffer Dharma. Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PloS one*, 12(5):e0178043, 2017.

[8] Yoshua Bengio. Time to rethink the publication process in machine learning. `https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/`, 2020. Accessed: 2020-9-20.

[9] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408*, 2020.

[10] Stella Biderman, Anima Anandkumar, Aylin Caliskan, Catherine D'Ignazio, and Ram Shankar Siva Kumar. Ai ethics and bias panel. The AI Village at DEF CON, 2020. URL `https://www.youtube.com/watch?v=7zswHvHR9cA`.

[11] Kevin W Bowyer, Michael King, and Walter Scheirer. The criminality from face illusion. *arXiv preprint arXiv:2006.03895*, 2020.

[12] L. Burke. Your interview with AI. *Inside Higher Ed*, 2019.

[13] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[14] Coalition for Critical Technology. Abolish the #techtoprisonpipeline. `https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16`, 2020. Accessed: 2020-9-20.

[15] Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. Mt-adapted datasheets for datasets: Template and repository. *arXiv preprint arXiv:2005.13156*, 2020.

[16] Deepjoy Das and Alok Chakrabarty. Human gait based gender identification system using hidden markov model and support vector machines. In *International Conference on Computing, Communication & Automation*, pages 268–272. IEEE, 2015.

[17] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[18] Trung Dung Do, Van Huan Nguyen, and Hakil Kim. Real-time and robust multiple-view gender classification using gait features in video surveillance. *Pattern Analysis and Applications*, 23(1): 399–413, 2020.

[19] Eyragon Eidam. The role of data analytics in predictive policing. *Government Technology*, 2016.

[20] Abdeljalil Gattal, Chawki Djeddi, Ameur Bensefia, and Abdellatif Ennaji. Handwriting based gender classification using cold and hinge features. In *International Conference on Image and Signal Processing*, pages 233–242. Springer, 2020.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *FAT ML*, 2018.

[22] Mahdi Hashemi and Margeret Hall. Criminal tendency detection from facial images and the gender bias effect. *Journal of Big Data*, 7(1):1–16, 2020.

[23] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, 2017.

[24] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

[25] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, March 2009.

[26] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[27] Deepak Kumar, Rajat Gupta, Ashirwad Sharma, and Sushil Kumar Saroj. Gender classification using skin patterns. In *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019.

[28] Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. In *Uncertainty in Artificial Intelligence*, pages 367–377. PMLR, 2020.

[29] Johnathan Littman. Waze hijacked l.a. in the name of convenience. can anyone put the genie back in the bottle? *Los Angeles Magazine*, 2019.

[30] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6): e271–e297, 2019.

[31] Steve Lopez. Column: How waze and google maps turned an encino neighborhood into a speedway. *Los Angeles Times*, 2018.

[32] Patrick McKenzie. Falsehoods programmers believe about names. *Kalzumeus Software*, 2010.

[33] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[34] T Den Ottelander, Arkadiy Dushatskiy, Marco Virgolin, and Peter AN Bosman. Local search is a remarkably strong baseline for neural architecture search. *arXiv preprint arXiv:2004.08996*, 2020.

[35] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9617–9626, 2019.

[36] Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol*, 5(11):e1000579, 2009.

[37] Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, 2014.

[38] Vinay Prabhu. https://twitter.com/vinayprabhu/status/1307460502017028096?s=20, 2020. Accessed: 2020-9-20.

[39] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[40] Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, pages 5485–5495, 2019.

[41] Filipe Rodrigues, Lourenco Mariana, Bernardete Ribeiro, and Francisco C. Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE T-PAMI*, 39(12): 2409 – 2422, 5 2017.

[42] W. Scheirer. How to make ai less racist. *Bulletin of the Atomic Scientists*, 2020.

[43] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. Baselines and a datasheet for the cerema awp dataset. *arXiv preprint arXiv:1806.04016*, 2018.

[44] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*, 2020.

[45] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: a guide for hci researchers. *interactions*, 26(4):62–65, 2019.

[46] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

[47] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53, 2020.

[48] A. Torralba, R. Fergus, and B. Freeman. Tiny Images Dataset Retraction. https://groups.csail.mit.edu/vision/TinyImages/, 2020. Accessed: 2020-9-20.

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

[50] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[51] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[52] Colin White, Sam Nolen, and Yash Savani. Local search is state of the art for nas benchmarks. *arXiv preprint arXiv:2005.02960*, 2020.

[53] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, pages 4038–4052, 2016.

[54] Jin Xie, Lei Zhang, Jane You, David Zhang, and Xiaofeng Qu. A study of hand back skin texture patterns for personal identification and gender classification. *Sensors*, 12(7):8691–8709, 2012.

[55] Saining Xie, Alexander Kirillov, Ross Girshick, and Kaiming He. Exploring randomly wired neural networks for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1284–1293, 2019.

[56] Amira E Youssef, Ahmed S Ibrahim, and A Lynn Abbott. Automated gender identification for arabic and english handwriting. In *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, pages 1–6. IET, 2013.

[57] Shiqi Yu, Tieniu Tan, Kaiqi Huang, Kui Jia, and Xinyu Wu. A study on gait-based gender classification. *IEEE Transactions on image processing*, 18(8):1905–1910, 2009.

[58] Kim Zetter. A famed hacker is grading thousands of programs — and may revolutionize software in the process. The Intercept, `https://rb.gy/igtlzz`, 2020. Accessed: 2020-9-20.

[59] Amy X. Zhang. `https://twitter.com/amyxzh/status/1307505876396158976?s=20`, 2020. Accessed: 2020-9-19.