The Values Encoded in Machine Learning Research

Abeba Birhane*
abeba@mozillafoundation.org
Mozilla Foundation & School of Computer Science, University College Dublin
Dublin, Ireland

Pratyusha Kalluri*
pkalluri@stanford.edu
Computer Science Department, Stanford University
Palo Alto, USA

Dallas Card*
dalc@umich.edu
School of Information, University of Michigan
Ann Arbor, USA

William Agnew*
wagnew3@cs.washington.edu
Paul G. Allen School of Computer Science and Engineering, University of Washington
Seattle, USA

Ravit Dotan*
ravit.dotan@berkeley.edu
Center for Philosophy of Science, University of Pittsburgh
Pittsburgh, USA

Michelle Bao*
baom@stanford.edu
Computer Science Department, Stanford University
Palo Alto, USA

ABSTRACT

Machine learning currently exerts an outsized influence on the world, increasingly affecting institutional practices and impacted communities. It is therefore critical that we question vague conceptions of the field as value-neutral or universally beneficial, and investigate what specific values the field is advancing. In this paper, we first introduce a method and annotation scheme for studying the values encoded in documents such as research papers. Applying the scheme, we analyze 100 highly cited machine learning papers published at premier machine learning conferences, ICML and NeurIPS. We annotate key features of papers which reveal their values: their justification for their choice of project, which attributes of their project they uplift, their consideration of potential negative consequences, and their institutional affiliations and funding sources. We find that few of the papers justify how their project connects to a societal need (15%) and far fewer discuss negative potential (1%). Through line-by-line content analysis, we identify 59 values that are uplifted in ML research, and, of these, we find that the papers most frequently justify and assess themselves based on Performance, Generalization, Quantitative evidence, Efficiency, Building on past work, and Novelty. We present extensive textual evidence and identify key themes in the definitions and operationalization of these values. Notably, we find systematic textual evidence that these top values are being defined and applied with assumptions and implications generally supporting the centralization of power. Finally, we find increasingly close ties between these highly cited papers and tech companies and elite universities.

KEYWORDS

Encoded values of ML, ICML, NeurIPS, Corporate ties, Power asymmetries

ACM Reference Format:
Abeba Birhane, Pratyusha Kalluri*, Dallas Card*, William Agnew*, Ravit Dotan*, and Michelle Bao*. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533083

1 INTRODUCTION

Over recent decades, machine learning (ML) has risen from a relatively obscure research area to an extremely influential discipline, actively being deployed in myriad applications and contexts around the world. Current discussions of ML frequently follow a historical strain of thinking which has tended to frame technology as "neutral", based on the notion that new technologies can be unpredictably applied for both beneficial and harmful purposes [64]. This claim of neutrality frequently serves as an insulation from critiques of AI and as permission to emphasize the benefits of AI [47, 58, 63], often without any acknowledgment that benefits and harms are distributed unevenly. Although it is rare to see anyone explicitly argue in print that ML is neutral, related ideas are part of contemporary conversation, including these canonical claims: long term impacts are too difficult to predict; sociological impacts are outside the expertise or purview of ML researchers [28]; critiques of AI are really misdirected critiques of those deploying AI with bad data ("garbage in, garbage out"), again outside the purview of many AI researchers; and proposals such as broader impact statements represent merely a "bureaucratic constraint" [3]. ML research is often cast as value-neutral and emphasis is placed on positive applications or potentials. Yet, the objectives and values of ML research are influenced by many social forces that shape factors including what research gets done and who benefits.[1] Therefore, it is important to challenge perceptions of neutrality and universal benefit, and document and understand the emergent values of the

*All authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533083

[1]For example, ML research is influenced by social factors including the personal preferences of researchers and reviewers, other work in science and engineering, the interests of academic institutions, funding agencies and companies, and larger systemic pressures, including systems of oppression.

field: what specifically the field is prioritizing and working toward. To this end, we perform an in-depth analysis of 100 highly cited NeurIPS and ICML papers from four recent years.

Our key contributions are as follows:

(1) **We present and open source a fine-grained annotation scheme for the study of values in documents such as research papers.**[2] To our knowledge, our annotation scheme is the first of its kind and opens the door to further qualitative and quantitative analyses of research. This is a timely methodological contribution, as institutions including prestigious ML venues and community organizations are increasingly seeking and reflexively conducting interdisciplinary study on social aspects of machine learning [6, 7, 12, 39].

(2) **We apply our scheme to annotate 100 influential ML research papers and extract their value commitments, including identifying 59 values significant in machine learning research.** These papers reflect and shape the values of the field. Like the annotation scheme, the resulting repository of over 3,500 annotated sentences is available and is valuable as foundation for further qualitative and quantitative study.

(3) **We perform extensive textual analysis to understand dominant values**: Performance, Generalization, Efficiency, Building on past work, and Novelty. Our analysis reveals that while these values may seem on their face to be purely technical, they are socially and politically charged: **we find systematic textual evidence corroborating that these values are currently defined and operationalized in ways that centralize power**, i.e., disproportionally benefit and empower the already powerful, while neglecting society's least advantaged.[3]

(4) **We present a quantitative analysis of the affiliations and funding sources of these influential papers. We find substantive and increasing presence of tech corporations.** For example, in 2008/09, 24% of these top cited papers had corporate affiliated authors, and in 2018/19 this statistic more than doubled, to 55%. Moreover, of these corporations connected to influential papers, the presence of "big-tech" firms, such as Google and Microsoft, more than tripled from 21% to 66%.

2 METHODOLOGY

To study the values of ML research, we conduct an in-depth analysis of ML research papers distinctively informative of these values.[4] We chose to focus on highly cited papers because they reflect and shape the values of the discipline, drawing from NeurIPS and ICML because they are the most prestigious of the long-running ML conferences.[5] Acceptance to these conferences is a valuable commodity used to evaluate researchers, and submitted papers are typically explicitly written so as to win the approval of the community, particularly the reviewers who will be drawn from that community. As such, these papers effectively reveal the values that authors believe are most valued by that community. Citations indicate amplification by the community, and help to position these papers as influential exemplars of ML research. To avoid detecting only short-lived trends, we drew papers from two recent years (2018/19[6]) and from ten years earlier (2008/09). We focused on conference papers because they tend to follow a standard format and allow limited space, meaning that researchers must make hard choices about what to emphasize. Collectively, an interdisciplinary team of researchers analyzed the 100 most highly cited papers from NeurIPS and ICML, from the years 2008, 2009, 2018, and 2019, annotating over 3,500 sentences drawn from them. In the context of expert content analysis, this constitutes a large scale annotation which allows us to meaningfully comment on central values.

Our team constructed an annotation scheme and applied it to manually annotate each paper, examining the abstract, introduction, discussion, and conclusion: (1) We examined the chain of reasoning by which each paper justified its contributions, which we call the *justificatory chain*, categorizing the extent to which papers used technical or societal problems to justify or motivate their contributions (Table 1).[7,8] (2) We carefully read each sentence of these sections line-by-line, inductively annotating any and all values uplifted by the sentence (Figure 1). We use a conceptualization of "value" that is widespread in philosophy of science in theorizing about values in sciences: a "value" of an entity is a property that is considered desirable for that kind of entity, e.g. regarded as a desirable attribute for machine learning research.[9] (3) We categorized the extent to which the paper included a discussion of potential negative impacts (Table 2).[8] (4) We documented and categorized the author affiliations and stated funding sources. In this paper, we provide complete annotations, quantize the annotations to quantify and present dominant patterns, and present randomly sampled excerpts and key themes in how these values become socially loaded.

To perform the line-by-line analysis and annotate the uplifted values (Figure 1), we used a hybrid inductive-deductive content

[2] We include our annotation scheme and all annotations at github.com/wagnew3/The-Values-Encoded-in-Machine-Learning-Research with a CC BY-NC-SA license.

[3] We understand this to be an interdisciplinary contribution: Scholarship on the values of ML (or alternatives) often faces dismissal based on perceived distance from prestigious ML research and quantifiable results. Meanwhile, philosophers of science have been working to understand the roles and political underpinnings of values in science for decades, e.g., in biology and social sciences [37, 42]. Our paper provides convincing qualitative and quantitative evidence of ML values and their political underpinnings, bridging ML research and both bodies of work.

[4] Because the aim of qualitative inquiry is depth of understanding, it is viewed as important to analyze information-rich documents (those that distinctively reflect and shape the central values of machine learning; for example, textual analysis of influential papers) in lieu of random sampling and broad analysis (for example, keyword frequencies in a large random sample of ML papers). This is referred to as the importance of purposive sampling [52].

[5] At the time of writing, NeurIPS and ICML, along with the newer conference ICLR, comprised the top 3 conferences according to h5-index (and h5-median) in the AI category on Google Scholar, by a large margin. Citation counts are based on the Semantic Scholar database.

[6] At the time of beginning annotation, 2018 and 2019 were the two most recent years available.

[7] In qualitative research, the term 'coding' is used to denote deductively categorizing text into selected categories as well as inductively annotating text with emergent categories. To avoid overloading computer science 'coding', we use the terms categorizing and annotating throughout this paper.

[8] We found the first three categories of this scheme were generally sufficient for our analysis. In service of rich understanding, we included the subtler fourth category. As much as possible, we steel-manned discussions: regardless of whether we were convinced or intrigued by a discussion, if it presented the level of detail typical when discussing projects' technical implications, then it was assigned category four.

[9] For example, speed can be described as valuable in an antelope [43]. Well-know scientific values include accuracy, consistency, scope, simplicity, and fruitfulness [37]. See [42] for a critical discussion of socially-laden aspects of these values in science.

analysis methodology and followed best practices [8, 29, 36, 44]: (i) We began with several values of interest based on prior literature, specifically seven ethical principles and user rights [5, 22, 31]. (ii) We randomly sampled a subset of 10 papers for initial annotation, reading sentence by sentence, deductively annotating for the values of interest and inductively adding new values as they emerged, by discussion until perfect consensus. The deductive component ensures we note and can speak to values of interest, and the inductive component enables discovery and impedes findings limited by bias or preconception by requiring textual grounding and focusing on emergent values [8, 36]. (iii) We annotated the full set of papers sentence by sentence. We followed the constant comparative method, in which we continually compared each text unit to the annotations and values list thus far, annotated for the values in the values list, held regular discussions, and we individually nominated and decided by consensus when sentences required inductively adding emergent values to the values list [23]. We used a number of established strategies in service of consistency which we discuss below. Following qualitative research best practices, we identified by consensus a small number of values we found were used synonymously or closely related and combined these categories, listing all merges in Appendix C.[10] (iv) In this paper, for each top value, we present randomly selected quotations of the value, richly describe the meaning of the value in context, present key themes in how the value is operationalized and becomes socially loaded, and illustrate its contingency by comparing to alternative values in the literature that might have been or might be valued instead.

We adhere to a number of best practices to establish reliability: We practice prolonged engagement, conducting long-term orientation to and analysis of data over more than a year (in lieu of short-term analysis that is dominated by preconceptions) [40]; We triangulate across researchers (six researchers) and points in time (four years) and place (two conferences) [17, 53]; We recode data coded early in the process [35]; We transparently publish the complete annotation scheme and all annotations [48]; We conduct negative case analysis, for example, drawing out and discussing papers with unusually strong connections to societal needs [40]; and we include a reflexivity statement in Appendix D describing our team in greater detail, striving to highlight relevant personal and disciplinary viewpoints.

The composition of our team confers additional validity to our work. We are a multi-racial, multi-gender team working closely, including undergraduate, graduate, and post-graduate researchers engaged with machine learning, NLP, robotics, cognitive science, critical theory, community organizing, and philosophy. This team captures several advantages: the nature of this team minimizes personal and intra-disciplinary biases, affords the unique combination of expertise required to read the values in complex ML papers, allows meaningful engagement with relevant work in other fields, and enabled best practices including continually clarifying the procedure, ensuring agreement, vetting consistency, reannotating, and discussing themes [36]. Across the annotating team, we found that annotators were able to make somewhat different and complementary inductive paper-level observations, while obtaining near or perfect consensus on corpus-level findings. To assess the consistency of paper-level annotations, 40% of the papers were double-annotated by paired annotators. During the inductive-deductive process of annotating sentences with values (ultimately annotating each sentence for the presence of 75 values), paired annotators agreed 87.0% of the time, and obtained a fuzzy Fleiss' kappa [34] on values per paper of 0.45, indicating moderate agreement. During the deductive process of categorizing the extent to which a paper included societal justification and negative potential impacts (ordinal categorization according to the schema in Table 1 and Table 2), paired annotators obtained substantial agreement, indicated by Fleiss' weighted kappa (κ=.60, κ=.79). Finally, at the corpus level we found substantial agreement: annotators identified the list of emergent values by perfect consensus, unanimously finding these values to be present in the papers. Across annotators, there was substantial agreement on the relative prevalence (ranking) of the values, indicated by Kendall's W [33] (W=.80), and we identified by consensus the five most dominant values, which we discuss in detail.

Manual analysis is necessary at all steps of the method (i-iv). Manual analysis is required for the central task of reading the papers and inductively identifying previously unobserved values. Additionally, once values have been established, we find manual analysis continues to be necessary for annotation. We find that many values are expressed in ways that are subtle, varied, or rely on contextual knowledge. We find current automated methods for labeling including keyword searches and basic classifiers miss new values, annotate poorly relative to manual annotation, and systematically skew the results towards values which are easy to identify, while missing or mischaracterizing values which are exhibited in more nuanced ways.[11] Accordingly, we find our use of qualitative methodology is indispensable. Reading all papers is key for contributing the textual analysis as well, as doing so includes developing a subtle understanding of how the values function in the text and understanding of taken for granted assumptions underlying the values.

In the context of an interdisciplinary readership, including ML and other STEM disciplines that foreground quantitative methodology, it is both a unique contribution and a limitation that this paper centers qualitative methodology. Ours is a significant and timely methodological contribution as there is rising interest in qualitatively studying the social values being encoded in ML, including reflexively by ML researchers [6, 7, 12, 39]. Simultaneously, the use of qualitative methodology in quantitative-leaning contexts could lead to misinterpretations. Human beliefs are complex and multitudinous, and it is well-established that when qualitative-leaning methodology is presented in quantitative-leaning contexts, it is possible for study of imprecise subject matter to be misinterpreted as imprecise study of subjects [10].

In brief, whereas quantitative analysis typically favors large random sampling and strict, statistical evidence in service of generalization of findings, qualitative analysis typically favors purposive sampling from information-rich context and richly descriptive evidence in service of depth of understanding [10, 44]. For both our

[10] For example, in Section 4.6, we discuss themes cutting across efficiency, sometimes referenced in the abstract and sometimes indicated by uplifting data efficiency, energy efficiency, fast, label efficiency, low cost, memory efficiency, or reduced training time.

[11] In Appendix E, we implement automatic annotation and empirically demonstrate these failure modes.

final list of values and specific annotation of individual sentences, different researchers might make somewhat different choices. However, given the overwhelming presence of certain values, the high agreement rate among annotators, and the similarity of observations made by our team, we believe other researchers following a similar approach would reach similar conclusions about what values are most frequently uplifted. Also, we cannot claim to have identified every relevant value in ML. Rather, we present a collection of such values; and by including important ethical values identified by past work, and specifically looking for these, we can confidently assert their relative absence in this set of papers. Finally, qualitative analysis is an effort to understand situations in their uniqueness, i.e., in this set of papers. Future work may determine whether and how to form conclusions about stratifications (e.g. between chosen years or conferences) and whether and how to use this qualitative analysis to construct new quantitative instruments to ascertain generalization (e.g. across more years or conferences) [20, 52]. Our study contributes unprecedent data and textual analysis and lays the groundwork for this future work.

3 QUANTITATIVE SUMMARY

In Figure 1, we plot the prevalence of values in 100 annotated papers. The top values are: performance (96% of papers), generalization (89%), building on past work (88%), quantitative evidence (85%), efficiency (84%), and novelty (77%). Values related to user rights and stated in ethical principles appeared very rarely if at all: none of the papers mentioned autonomy, justice, or respect for persons. In Table 1, we show the distribution of justification scores. Most papers only justify how they achieve their internal, technical goal; 68% make no mention of societal need or impact, and only 4% make a rigorous attempt to present links connecting their research to societal needs. In Table 2, we show the distribution of negative impact discussion scores. One annotated paper included a discussion of negative impacts and a second mentioned the possibility of negative impacts. 98% of papers contained no reference to potential negative impacts. In Figure 3, we show stated connections (funding ties and author affiliations) to institutions. Comparing papers written in 2008/2009 to those written in 2018/2019, ties to corporations nearly doubled to 79% of all annotated papers, ties to big tech more than tripled, to 66%, while ties to universities declined to 81%, putting the presence of corporations nearly on par with universities. In the next section, we present extensive qualitative examples and analysis of our findings.

4 TEXTUAL ANALYSIS

4.1 Justifications

We find papers typically justify their choice of project by contextualizing it within a broader goal and giving a chain of justification from the broader goal to the particular project pursued in the paper. These justifications reveal priorities:

Papers typically motivate their projects by appealing to the needs of the ML research community and rarely mention potential societal benefits. Research-driven needs of the ML community include researcher understanding (e.g., understanding the effect of pre-training on performance/robustness, theoretically understanding multi-layer networks) as well as more practical research problems (e.g., improving efficiency of models for large datasets, creating a new benchmark for NLP tasks).

Even when societal needs are mentioned as part of the justification of the project, the connection is loose. Some papers do appeal to needs of broader society, such as building models with realistic assumptions, catering to more languages, or "understanding the world". Yet almost no papers explain how their project promotes a social need they identify by giving the kind of rigorous justification that is typically expected of and given for technical contributions.

The cursory nature of the connection between societal needs and the content of the paper also manifests in the fact that the societal needs, or the applicability to the real world, is often only discussed in the beginning of the papers. From papers that mention applicability to the real world, the vast majority of mentions are in the Introduction section, and applicability is rarely engaged with afterwards. Papers tend to introduce the problem as useful for applications in object detection or text classification, for example, but rarely justify why an application is worth contributing to, or revisit how they particularly contribute to an application as their result.

4.2 Discussion of Negative Potential

Although a plethora of work exists on sources of harm that can arise in relation to ML research [6, 14, 24, 27, 59], we observe that these discussions are ignored in these influential conference publications.

It is extremely rare for papers to mention negative potential at all. Just as the goals of the papers are largely inward-looking, prioritizing the needs of the ML research community, these papers fail to acknowledge both broader societal needs and societal impacts. This norm is taken for granted: none of these papers offer any explanation for why they cannot speak to negative impacts. These observations correspond to a larger trend in the ML research community of neglecting to discuss aspects of the work that are not strictly positive.

The lack of discussion of potential harms is especially striking for papers which deal with contentious application areas, such as surveillance and misinformation. These include papers, for example, that advance identification of people in images, face-swapping, and video synthesis. These papers contain no mention of the well-studied negative potential of facial surveillance, DeepFakes, or misleading videos.

Among the two papers that do mention negative potential, the discussions were mostly abstract and hypothetical, rather than grounded in the concrete negative potential of their specific contributions. For example, authors may acknowledge "possible unwanted social biases" when applying models to a real-world setting, without commenting on let alone assessing the social biases encoded in the authors' proposed model.

4.3 Stated values

The dominant values that emerged from the annotated corpus are: Performance, Generalization, Building on past work, Quantitative evidence, Efficiency, and Novelty. These are often portrayed as innate and purely technical. However, the following analysis of these values shows how they can become politically loaded in the

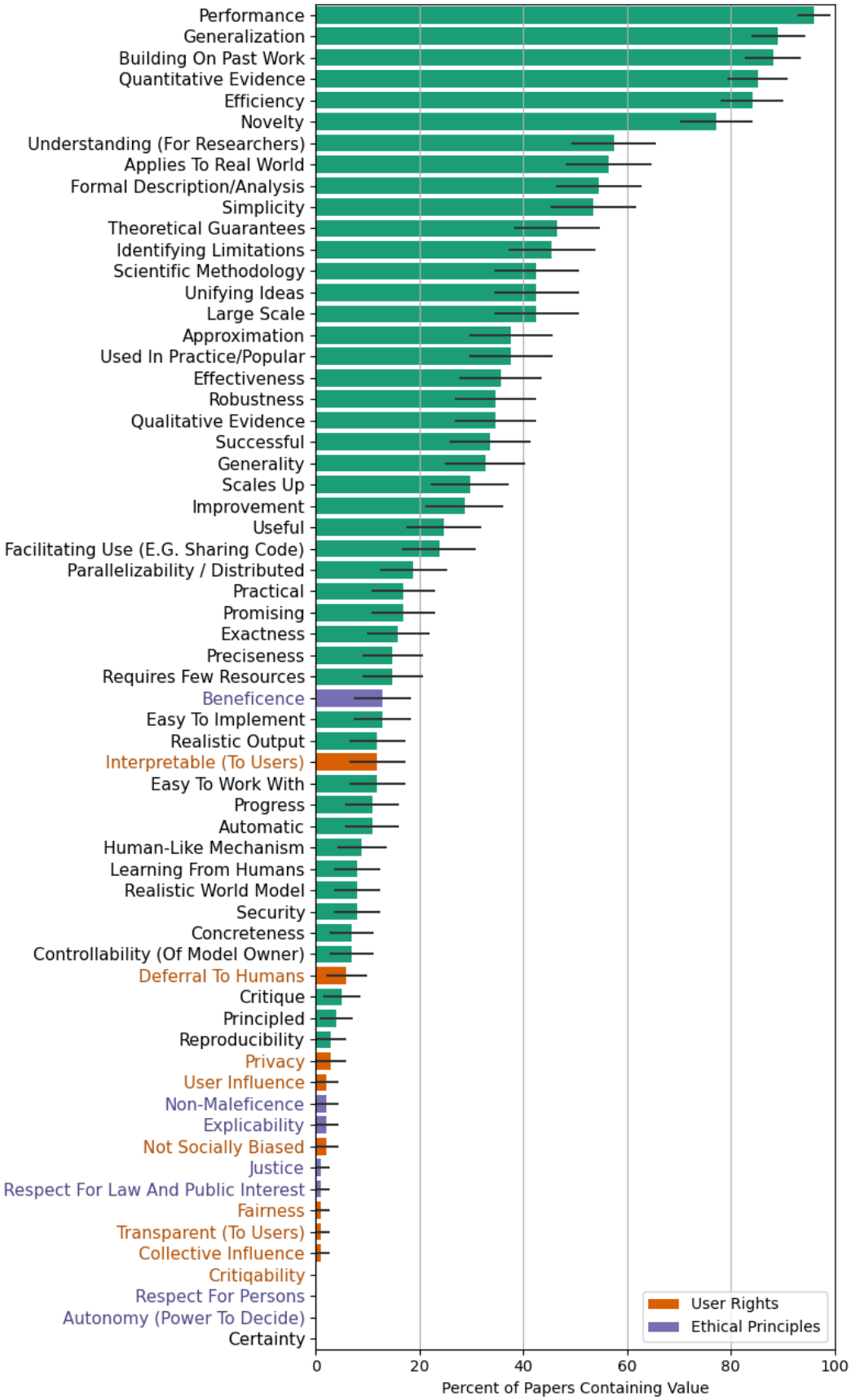

Figure 1: Proportion of annotated papers that uplift each value.

Table 1: Annotations of justificatory chain.

Justificatory Chain	% of Papers
Does not mention societal need	68%
States but does not justify how it connects to a societal need	17%
States and somewhat justifies how it connects to a societal need	11%
States and rigorously justifies how it connects to a a societal need	4%

Table 2: Annotations of discussed negative potential.

Discussion of Negative Potential	% of Papers
Does not mention negative potential	98%
Mentions but does not discuss negative potential	1%
Discusses negative potential	1%
Deepens our understanding of negative potential	0%

process of prioritizing and operationalizing them: sensitivity to the way that they are operationalized, and to the fact that they are uplifted at all, reveals value-laden assumptions that are often taken for granted. To provide a sense of what the values look like in context, Tables 3, 4, 5 and 6 present randomly selected examples of sentences annotated with the values of Performance, Generalization, Efficiency, Building on past work, and Novelty respectively. Extensive additional examples can be found in Appendix H.[12] For each of these prominent values, we quantify its dominance, identify constituent values that contribute to this value, challenge a conception of the value as politically neutral, identify key themes in how the value is socialy loaded, and we cite alternatives to its dominant conceptualization that may be equally or more valid, interesting, or socially beneficial. When values seem neutral or innate, we have encouraged ourselves, and now encourage the reader, to remember that values once held to be intrinsic, obvious, or definitional have in many cases been found harmful and transformed over time and purportedly neutral values warrant careful consideration.

4.4 Performance

Emphasizing performance is the most common way by which papers attempt to communicate their contributions, by showing a specific, quantitative, improvement over past work, according to some metric on a new or established dataset. For some reviewers, obtaining better performance than any other system—a "state-of-the-art" (SOTA) result—is seen as a noteworthy, or even necessary, contribution [57].

Despite acknowledged issues with this kind of evaluation (including the artificiality of many datasets, and the privileging of "tricks" over insight; 21, 41), performance is typically presented as intrinsic to the field. Frequently, the value of Performance is indicated by specifically uplifting accuracy or state of the art results, which are presented as similarly intrinsic. However, models are not simply "well-performing" or "accurate" in the abstract but always in relation to and as quantified by some *metric* on some *dataset*. Examining definition and operationalization of performance values, we identify three key social aspects.

Table 3: Random examples of *performance*, the most common emergent value.

"Our model significantly outperforms SVM's, and it also outperforms convolutional neural nets when given additional unlabeled data produced by small translations of the training images."
"We show in simulations on synthetic examples and on the IEDB MHC-I binding dataset, that our approach outperforms well-known convex methods for multi-task learning, as well as related non-convex methods dedicated to the same problem."
"Furthermore, the learning accuracy and performance of our LGP approach will be compared with other important standard methods in Section 4, e.g., LWPR [8], standard GPR [1], sparse online Gaussian process regression (OGP) [5] and υ-support vector regression (υ-SVR) [11], respectively."
"In addition to having theoretically sound grounds, the proposed method also outperformed state-of-the-art methods in two experiments with real data."
"We prove that unlabeled data bridges this gap: a simple semisupervised learning procedure (self-training) achieves high robust accuracy using the same number of labels required for achieving high standard accuracy."
"Experiments show that PointCNN achieves on par or better performance than state-of-the-art methods on multiple challenging benchmark datasets and tasks."
"Despite its impressive empirical performance, NAS is computationally expensive and time consuming, e.g. Zoph et al. (2018) use 450 GPUs for 3-4 days (i.e. 32,400-43,200 GPU hours)."
"However, it is worth examining why this combination of priors results in superior performance."
"In comparisons with a number of prior HRL methods, we find that our approach substantially outperforms previous state-of-the-art techniques."
"Our proposed method addresses these issues, and greatly outperforms the current state of the art."

[12] To avoid the impression that we are mainly interested in drawing attention to specific papers, we omit attribution for individual examples, but include a list of all annotated papers in Appendix I. Note that most sentences are annotated with multiple values; for example, there can be overlap in sentences annotated with *performance* and sentences annotated with *generalization*.

✦ **Performance values are consistently and without discussion operationalized as correctness averaged across individual predictions, giving equal weight to each instance.** However, choosing equal weights when averaging is a value-laden move which might deprioritize those underrepresentated in the data or the world, as well as societal and evaluee needs and preferences regarding inclusion. Extensive research in ML fairness and related fields has considered alternatives, but we found no such discussions among the influential papers we examined.

✦ **Datasets are typically preestablished, large corpora with discrete "ground truth" labels.** They are often driven purely by past work, so as to demonstrate improvement over a previous baseline (see also §4.7). Another common justification for using a certain dataset is claimed applicability to the "real world". Assumptions about how to characterize the "real world" are value-laden. One preestablished and typically perpetuated assumption is the availability of very large datasets. However, presupposing the availability of large datasets is non-neutral and power centralizing because it encodes favoritism to those with resources to obtain and process them [19]. Additionally, the welfare, consent, or awareness of the datafied subjects whose images end up in a large scale image dataset, for example, are not considered in the annotated papers. Further overlooked assumptions include that the real world is binary or discrete, and that datasets come with a predefined ground-truth label for each example, presuming that a true label always exists "out there" independent of those carving it out, defining and labelling it. This contrasts against marginalized scholars' calls for ML models that allow for non-binaries, plural truths, contextual truths, and many ways of being [15, 25, 38].

✦ **The prioritization of performance values is so entrenched in the field that generic success terms, such as "success", "progress", or "improvement" are used as synonyms for performance and accuracy.** However, one might alternatively invoke generic success to mean increasingly safe, consensual, or participatory ML that reckons with impacted communities and the environment. In fact, "performance" itself is a general success term that could have been associated with properties other than accuracy and SOTA.

Table 4: Random examples of *generalization*, the second most common emergent value.

"The range of applications that come with generative models are vast, where audio synthesis [55] and semi-supervised classification [38, 31, 44] are examples hereof."
"Furthermore, the infinite limit could conceivably make sense in deep learning, since over-parametrization seems to help optimization a lot and doesn't hurt generalization much [Zhang et al., 2017]: deep neural nets with millions of parameters work well even for datasets with 50k training examples."
"Combining the optimization and generalization results, we uncover a broad class of learnable functions, including linear functions, two-layer neural networks with polynomial activation $\phi(z) = z^{2l}$ or cosine activation, etc."
"We can apply the proposed method to solve regularized least square problems, which have the loss function $(1 - y_i \omega^T x_i)^2$ in (1)."
"The result is a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets."
"Our proposed invariance measure is broadly applicable to evaluating many deep learning algorithms for many tasks, but the present paper will focus on two different algorithms applied to computer vision."
"We show how both multitask learning and semi-supervised learning improve the generalization of the shared tasks, resulting in state-of-the-art performance."
"We have also demonstrated that the proposed model is able to generalize much better than LDA in terms of both the log-probability on held-out documents and the retrieval accuracy."
"We define a rather general convolutional network architecture and describe its application to many well known NLP tasks including part-of-speech tagging, chunking, named-entity recognition, learning a language modeland the task of semantic role-labeling"
"We demonstrate our algorithm on multiple datasets and show that it outperforms relevant baselines."

4.5 Generalization

We observe that a common way of appraising the merits of one's work is to claim that it generalizes well. Notably, generalization is understood in terms of the dominant value, performance: a model is perceived as generalizing when it achieves good performance on a range of samples, datasets, domains, tasks, or applications. In fact, the value of generalization is sometimes indicated by referencing generalization in the abstract and other times indicated by specifically uplifting values such as Minimal discrepancy between train/test samples or Flexibility/extensibility, e.g., to other tasks. We identify three key socially loaded aspects of how generalization is defined and operationalized.

✦ **Only certain datasets, domains, or applications are valued as indicators of model generalization.** Typically, a paper shows that a model generalizes by showing that it performs well on multiple tasks or datasets. However, like the tasks and datasets indicating performance, the choice of particular tasks and datasets indicating generalization is rarely justified; the choice of tasks can often seem arbitrary, and authors often claim generalization while rarely presenting discussion or analysis indicating their results will generalize outside the carefully selected datasets, domains or applications, or to more realistic settings, or help to directly address societal needs.

✦ **Prizing generalization leads institutions to harvest datasets from various domains, and to treat these as the only datasets that matter in the space of problems**. Papers prizing generalization implicitly and sometimes explicitly prioritize reducing every scenario top-down to a common set of representations or affordances, rather than treating each setting as meaningfully unique and potentially motivating technologies or lack thereof that are fundamentally different from the current standard. Despite vague associations between generalization and accessible technology for diverse peoples, in practice work on generalization frequently targets one model to rule them all, denigrating diverse access needs. Critical scholars have advocated for valuing *context*,

which may stand opposed to striving for generalization [18]. Others have argued that this kind of totalizing lens (in which model developers have unlimited power to determine how the world is represented) leads to *representational* harms, due to applying a single representational framework to everything [1, 16].

✦ **The belief that generalization is possible assumes new data will be or should be treated similarly to previously seen data.** When used in the context of ML, the assumption that the future resembles the past is often problematic as past societal stereotypes and injustice can be encoded in the process [50]. Furthermore, to the extent that predictions are performative [54], especially predictions that are enacted, those ML models which are deployed to the world will contribute to shaping social patterns. None of the annotated papers attempt to counteract this quality or acknowledge its presence.

4.6 Efficiency

In the annotated papers, we find that saying that a model is efficient typically indicates the model uses less of some resource, e.g., data efficiency, energy efficiency, label efficiency, memory efficiency, being low cost, fast, or having reduced training time. We find that the definition and operationalization of efficiency encodes key social priorities, namely *which kind of efficiency matters* and *to what end*.

✦ **Efficiency is commonly referenced to indicate the ability to scale up, not to save resources.** For example, a more efficient inference method allows you to do inference in much larger models or on larger datasets, using the same amount of resources used previously, or more. This mirrors the classic Jevon's paradox: greater resource efficiency often leads to overall greater utilization of that resource. This is reflected in our value annotations, where 84% of papers mention valuing efficiency, but only 15% of those value requiring *few resources*. When referencing the consequences of efficiency, many papers present evidence that efficiency enables scaling up, while none of the papers present evidence that efficiency can facilitate work by low-resource communities or can lessen resource extraction – e.g. less hardware or data harvesting or lower carbon emissions. In this way, valuing efficiency facilitates and encourages the most powerful actors to scale up their computation to ever higher orders of magnitude, making their models even less accessible to those without resources to use them and decreasing the ability to compete with them. Alternative usages of efficiency could encode accessibility instead of scalability, aiming to create more equitable conditions.

4.7 Novelty and Building on Past Work

Most authors devote space in the introduction to positioning their paper in relation to past work, and describing what is novel. Building on past work is sometimes referenced broadly and other times is indicated more specifically as building on classic work or building on recent work. In general, mentioning past work serves to signal awareness of related publications, to establish the new work as relevant to the community, and to provide the basis upon which to make claims about what is new. Novelty is sometimes suggested implicitly (e.g., "we develop" or "we propose"), but frequently it is emphasized explicitly (e.g. "a new algorithm" or "a novel approach"). The emphasis on novelty is common across many academic fields [60, 61]. The combined focus on novelty and building on past work establishes a continuity of ideas, and might be expected to contribute to the self-correcting nature of science [45]. However, this is not always the case [30] and attention to the ways novelty and building on past work are defined and implemented reveals two key social commitments.

Table 5: Random examples of *efficiency*, the fifth most common emergent value.

"Our model allows for controllable yet efficient generation of an entire news article – not just the body, but also the title, news source, publication date, and author list."
"We show that Bayesian PMF models can be efficiently trained using Markov chain Monte Carlo methods by applying them to the Netflix dataset, which consists of over 100 million movie ratings."
"In particular, our EfficientNet-B7 surpasses the best existing GPipe accuracy (Huang et al., 2018), but using 8.4x fewer parameters and running 6.1x faster on inference."
"Our method improves over both online and batch methods and learns faster on a dozen NLP datasets."
"We describe efficient algorithms for projecting a vector onto the $\ell 1$-ball."
"Approximation of this prior structure through simple, efficient hyperparameter optimization steps is sufficient to achieve these performance gains."
"We have developed a new distributed agent IMPALA (Importance Weighted Actor-Learner Architecture) that not only uses resources more efficiently in single-machine training but also scales to thousands of machines without sacrificing data efficiency or resource utilisation."
"In this paper we propose a simple and efficient algorithm SVP (Singular Value Projection) based on the projected gradient algorithm"
"We give an exact and efficient dynamic programming algorithm to compute CNTKs for ReLU activation."
"In contrast, our proposed algorithm has strong bounds, requires no extra work for enforcing positive definiteness, and can be implemented efficiently."

✦ **Technical novelty is most highly valued.** The highly-cited papers we examined mostly tend to emphasize the novelty of their proposed method or of their theoretical result. Very few uplifted their paper on the basis of applying an existing method to a novel domain, or for providing a novel philosophical argument or synthesis. We find a clear emphasis on technical novelty, rather than critique of past work, or demonstration of measurable progress on societal problems, as has previously been observed [62].

✦ **Although introductions sometimes point out limitations of past work so as to further emphasize the contributions of their own paper, they are rarely explicitly critical of other papers in terms of datasets, methods, or goals.** Indeed, papers uncritically reuse the same datasets for years or decades to benchmark their algorithms, even if those datasets fail to represent more realistic contexts in which their algorithms will be used [6]. Novelty is denied to work that critiques or rectifies socially harmful aspects of existing datasets and goals, and this occurs in tandem

Table 6: Random examples of *building on past work* and *novelty*, the third and sixth most common emergent values, respectively.

Building on past work
"Recent work points towards sample complexity as a possible reason for the small gains in robustness: Schmidt et al. [41] show that in a simple model, learning a classifier with non-trivial adversarially robust accuracy requires substantially more samples than achieving good 'standard' accuracy."
"Experiments indicate that our method is much faster than state of the art solvers such as Pegasos, TRON, SVMperf, and a recent primal coordinate descent implementation."
"There is a large literature on GP (response surface) optimization."
"In a recent breakthrough, Recht et al. [24] gave the first nontrivial results for the problem obtaining guaranteed rank minimization for affine transformations A that satisfy a restricted isometry property (RIP)."
"In this paper, we combine the basic idea behind both approaches, i.e., LWPR and GPR, attempting to get as close as possible to the speed of local learning while having a comparable accuracy to Gaussian process regression"
Novelty
"In this paper, we propose a video-to-video synthesis approach under the generative adversarial learning framework."
"Third, we propose a novel method for the listwise approach, which we call ListMLE."
"The distinguishing feature of our work is the use of Markov chain Monte Carlo (MCMC) methods for approximate inference in this model."
"To our knowledge, this is the first attack algorithm proposed for this threat model."
"Here, we focus on a different type of structure, namely output sparsity, which is not addressed in previous work."

Figure 2: Corporate and Big Tech author affiliations. The percent of papers with Big Tech author affiliations increased from 13% in 2008/09 to 47% in 2018/19.

with strong pressure to benchmark on them and thereby perpetuate their use, enforcing a conservative bent to ML research.

5 CORPORATE AFFILIATIONS AND FUNDING

Quantitative summary. Our analysis shows substantive and increasing corporate presence in the most highly-cited papers. In 2008/09, 24% of the top cited papers had *corporate affiliated authors*, and in 2018/19 this statistic more than doubled to 55%. Furthermore, we also find a much greater concentration of a few large tech firms, such as Google and Microsoft, with the presence of these "big tech" firms (as identified in [4]) increasing nearly fourfold, from 13% to 47% (Figure 2). The fraction of the annotated papers with corporate ties by *corporate affiliated authors or corporate funding* dramatically increased from 45% in 2008/09 to 79% in 2018/19 (Figure 3). These findings are consistent with contemporary work indicating a pronounced corporate presence in ML research: in an automated analysis of peer-reviewed papers from 57 major computer science conferences, Ahmed and Wahed [4] show that the share of papers with corporate affiliated authors increased from 10% in 2005 for both ICML and NeurIPS to 30% and 35% respectively in 2019. Our analysis shows that corporate presence is even more pronounced in those papers from ICML and NeurIPS that end up receiving the most citations. In addition, we found paramount domination of elite universities in our analysis as shown in Figure 3. Of the total papers with university affiliations, we found 80% were from elite universities (defined as the top 50 universities by QS World University Rankings, following past work [4]).

Analysis. The influence of powerful players in ML research is consistent with field-wide value commitments that centralize power. Others have argued for causal connections. For example, Abdalla and Abdalla [2] argue that big tech sway and influence academic and public discourse using strategies which closely resemble strategies used by Big Tabacco. Moreover, examining the prevalent values of big tech, critiques have repeatedly pointed out that objectives such as efficiency, scale, and wealth accumulation [26, 50, 51] drive the industry at large, often at the expense of individuals rights, respect for persons, consideration of negative impacts, beneficence, and justice. Thus, the top stated values of ML that we presented in this paper such as performance, generalization, and efficiency may not only enable and facilitate the realization of big tech's objectives, but also suppress values such as beneficence, justice, and inclusion. A "state-of-the-art" large image dataset, for example, is instrumental for large scale models, further benefiting ML researchers and big tech in possession of huge computing power. In the current climate — where values such as accuracy, efficiency, and scale, as currently defined, are a priority, and there is a pattern of centralization of power — user safety, informed consent, or participation may be perceived as costly and time consuming, evading social needs.

6 DISCUSSION AND RELATED WORK

There is a foundational understanding in Science, Technology, and Society Studies (STS), Critical Theory, and Philosophy of Science that science and technologies are inherently value-laden, and these values are encoded in technological artifacts, many times in contrast to a field's formal research criteria, espoused consequences, or ethics guidelines [9, 13, 65]. There is a long tradition of exposing and critiquing such values in technology and computer science. For example, Winner [65] introduced several ways technology can encode political values. This work is closely related to Rogaway [56], who notes that cryptography has political and moral dimensions and argues for a cryptography that better addresses societal needs.

Our paper extends these critiques to the field of ML. It is a part of a rich space of interdisciplinary critiques and alternative lenses used to examine the field. Works such as [11, 46] critique AI, ML, and data using a decolonial lens, noting how these technologies

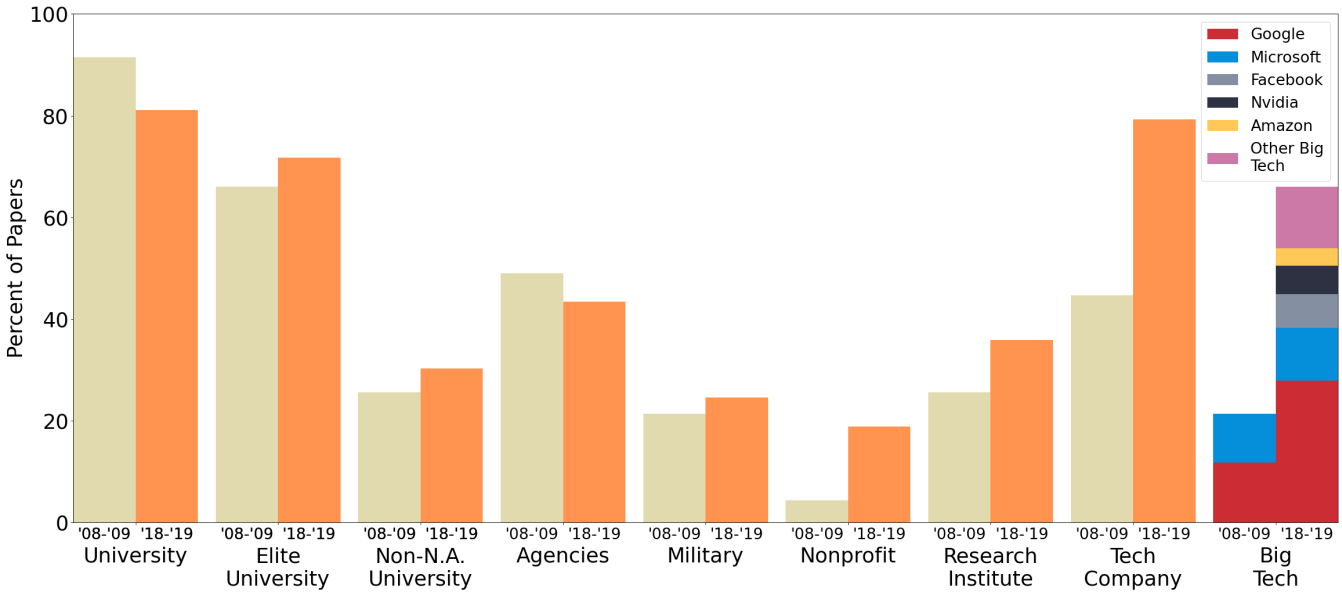

Figure 3: Affiliations and funding ties.
From 2008/09 to 2018/19, the percent of papers tied to nonprofits, research institutes, and tech companies increased substantially. Most significantly, ties to Big Tech increased threefold and overall ties to tech companies increased to 79%. Non-N.A. Universities are those outside the U.S. and Canada.

replicate colonial power relationships and values, and propose decolonial values and methods. Others [9, 18, 49] examine technology and data science from an anti-racist and intersectional feminist lens, discussing how our infrastructure has largely been built by and for white men; D'Ignazio and Klein [18] present a set of alternative principles and methodologies for an intersectional feminist data science. Similarly, Kalluri [32] denotes that the core values of ML are closely aligned with the values of the most privileged and outlines a vision where ML models are used to shift power from the most to the least powerful. Dotan and Milli [19] argue that the rise of deep learning is value-laden, promoting the centralization of power among other political values. Many researchers, as well as organizations such as Data for Black Lives, the Algorithmic Justice League, Our Data Bodies, the Radical AI Network, Indigenous AI, Black in AI, and Queer in AI, explicitly work on continuing to uncover particular ways technology in general and ML in particular can encode and amplify racist, sexist, queerphobic, transphobic, and otherwise marginalizing values, while simultaneously working to actualize alternatives [14, 55].

There has been considerable growth over the past few years in institutional, academic, and grassroots interest in the societal impacts of ML, as reflected in the rise of relevant grassroots and non-profit organizations, the organizing of new workshops, the emergence of new conferences such as FAccT, and changes to community norms, such as the required broader impacts statements at NeurIPS. We present this paper in part to make visible the present state of the field and to demonstrate its contingent nature; it could be otherwise. For individuals, communities, and institutions wading through difficult-to-pin-down values of the field, as well as those striving toward alternative values, it is advantageous to have a characterization of the way the field is now — to serve as both a confirmation and a map for understanding, shaping, dismantling, or transforming what is, and for articulating and bringing about alternative visions.

7 CONCLUSION

In this study, we find robust evidence against the vague conceptualization of the discipline of ML as value-neutral. Instead, we investigate the ways that the discipline of ML is inherently value-laden. Our analysis of highly influential papers in the discipline finds that they not only favor the needs of research communities and large firms over broader social needs, but also that they take this favoritism for granted, not acknowledging critiques or alternatives. The favoritism manifests in the choice of projects, the lack of consideration of potential negative impacts, and the prioritization and operationalization of values such as performance, generalization, efficiency, and novelty. These values are operationalized in ways that disfavor societal needs. Moreover, we uncover an overwhelming and increasing presence of big tech and elite universities in these highly cited papers, which is consistent with a system of power-centralizing value-commitments. The upshot is that the discipline of ML is not value-neutral. We present extensive quantitative and qualitative evidence that it is socially and politically loaded, frequently neglecting societal needs and harms, while prioritizing and promoting the concentration of resources, tools, knowledge, and power in the hands of already powerful actors.

ACKNOWLEDGMENTS

We would like to thank Luke Stark, Dan Jurafsky, and Sarah K. Dreier for helpful feedback on this work. We owe gratitude and

accountability to the long history of work exposing how technology shifts power, work primarily done by communities at the margins. Abeba Birhane was supported in part by Science Foundation Ireland grant 13/RC/2094_2. Pratyusha Kalluri was supported in part by the Open Phil AI Fellowship. Dallas Card was supported in part by the Stanford Data Science Institute. William Agnew was supported by an NDSEG Fellowship.

REFERENCES

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in Representation: Quantifying Stereotyping as a Representational Harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*.

[2] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462563

[3] Grace Abuhamad and Claudel Rheault. 2020. Like a Researcher Stating Broader Impact For the Very First Time. *arXiv preprint arXiv:2011.13032* (2020).

[4] Nur Ahmed and Muntasir Wahed. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *arXiv preprint arXiv:2010.15581* (2020).

[5] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research.* Technical Report. U.S. Department of Homeland Security.

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of FAccT* (2021).

[7] Samy Bengio and Deborah Raji. 2021. A Retrospective on the NeurIPS 2021 Ethics Review Process. https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/

[8] Mariette Bengtsson. 2016. How to plan and perform a qualitative study using content analysis. *NursingPlus Open* 2 (2016), 8–14.

[9] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Wiley.

[10] Bruce L. Berg and Howard Lune. 2017. *Qualitative research methods for the social sciences* (ninth edition ed.). Pearson.

[11] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *SCRIPTed* 17, 2 (2020).

[12] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.485

[13] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences.* MIT press.

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*.

[15] Sasha Costanza-Chock. 2018. Design Justice, AI, and Escape from the Matrix of Domination. *Journal of Design and Science* (2018).

[16] Kate Crawford. 2017. The Trouble with Bias. (2017). NeurIPS Keynote.

[17] Norman K Denzin. 2017. *Sociological methods: a sourcebook.* McGraw-Hill.

[18] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism.* MIT Press.

[19] Ravit Dotan and Smitha Milli. 2019. Value-Laden Disciplinary Shifts in Machine Learning. *arXiv preprint arXiv:1912.01172* (2019).

[20] Louise Doyle, Catherine McCabe, Brian Keogh, Annemarie Brady, and Margaret McCann. 2020. An overview of the qualitative descriptive design within nursing research. *Journal of Research in Nursing* 25, 5 (Aug 2020), 443–455. https://doi.org/10.1177/1744987119880234

[21] Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of EMNLP*.

[22] Luciano Floridi and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1, 1 (2019).

[23] Barney G. Glaser and Anselm L. Strauss. 1999. *The discovery of grounded theory: strategies for grounded research.* Aldine de Gruyter.

[24] Ben Green. 2019. 'Good' isn't Good Enough. In *NeurIPS Joint Workshop on AI for Social Good*.

[25] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings CHI*.

[26] Alex Hanna and Tina M. Park. 2020. Against Scale: Provocations and Resistances to Scale Thinking. *arXiv preprint arXiv:2010.08850* (2020).

[27] Kashmir Hill. 2020. Wrongfully Accused by an Algorithm. *The New York Times* (2020). https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[28] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of CHI*.

[29] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9 (2005), 1277–1288. https://doi.org/10.1177/1049732305276687

[30] John P. A. Ioannidis. 2012. Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science* 7, 6 (2012), 645–654.

[31] Pratyusha Kalluri. 2019. The Values of Machine Learning. https://slideslive.com/38923453/the-values-of-machine-learning https://slideslive.com/38923453/the-values-of-machine-learning.

[32] Pratyusha Kalluri. 2020. Don't ask if Artificial Intelligence is Good or Fair, ask how it Shifts Power. *Nature* 583, 7815 (2020), 169–169.

[33] Maurice G Kendall and B Babington Smith. 1939. The problem of m rankings. *The annals of mathematical statistics* 10, 3 (1939), 275–287.

[34] Andrei P Kirilenko and Svetlana Stepchenkova. 2016. Inter-coder agreement in one-to-many classification: fuzzy kappa. *PloS one* 11, 3 (2016), e0149787.

[35] Laura Krefting. 1991. Rigor in Qualitative Research: The Assessment of Trustworthiness. *American Journal of Occupational Therapy* 45, 3 (03 1991), 214–222. https://doi.org/10.5014/ajot.45.3.214

[36] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology.* Sage Publications.

[37] Thomas S. Kuhn. 1977. Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change.* University of Chicago Press, 320–39.

[38] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, et al. 2020. Indigenous Protocol and Artificial Intelligence Position Paper. (2020).

[39] T. Lewis, S. P. Gangadharan, M. Saba, and T. Petty. 2018. *Digital Defense Playbook: Community power tools for reclaiming data.* Our Data Bodies.

[40] Yvonna S. Lincoln and Egon G. Guba. 2006. *Naturalistic inquiry.* Sage Publ.

[41] Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling Trends in Machine Learning Scholarship: Some ML Papers Suffer from Flaws That Could Mislead the Public and Stymie Future Research. *Queue* 17, 1 (2019), 45–77. https://doi.org/10.1145/3317287.3328534

[42] Helen E. Longino. 1996. Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In *Feminism, Science, and the Philosophy of Science*, Lynn Hankinson Nelson and Jack Nelson (Eds.). Springer Netherlands, 39–58.

[43] Ernan McMullin. 1982. Values in science. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association.*

[44] Sharan B Merriam and Robin S Grenier. 2019. *Qualitative Research in Practice.* Jossey-Bass.

[45] Robert K. Merton. 1973. *The Sociology of Science: Theoretical and Empirical Investigations.* University of Chicago press.

[46] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Rheory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* 33 (2020), 659––684.

[47] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462608

[48] Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. *Evidence Based Nursing* 18, 2 (Apr 2015), 34–35.

[49] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[50] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books.

[51] Frank Pasquale. 2015. *The black box society.* Harvard University Press.

[52] Michael Quinn Patton. 1990. *Qualitative Evaluation and Research Methods.* Sage.

[53] M Q Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research* 34, 5 (Dec 1999).

[54] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of ICML*.

[55] Vinay Uday Prabhu and Abeba Birhane. 2020. Large Image Datasets: A Pyrrhic Win for Computer Vision? *arXiv preprint arXiv:2006.16923* (2020). https://arxiv.org/abs/2006.16923

[56] Phillip Rogaway. 2015. The Moral Character of Cryptographic Work. Cryptology ePrint Archive, Report 2015/1162. https://eprint.iacr.org/2015/1162.

[57] Anna Rogers. 2019. Peer review in NLP: reject-if-not-SOTA. *Hacking Semantics blog* (2019). https://hackingsemantics.xyz/2020/reviewing-models/#everything-wrong-with-reject-if-not-sota

[58] Daniela Rus. 2018. Rise of the robots: Are you ready? *Financial Times Magazine* (March 2018). https://www.ft.com/content/e31c4986-20d0-11e8-a895-1ba1f72c2c11

[59] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002* (2019). http://arxiv.org/abs/1901.10002

[60] Denis Trapido. 2015. How Novelty in Knowledge Earns Recognition: The Role of Consistent Identities. *Research Policy* 44, 8 (2015), 1488–1500.
[61] Christiaan H. Vinkers, Joeri K. Tijdink, and Willem M. Otte. 2015. Use of Positive and Negative Words in Scientific PubMed Abstracts between 1974 and 2014: Retrospective Analysis. *BMJ* 351 (2015).
[62] Kiri Wagstaff. 2012. Machine Learning that Matters. In *Proceedings of ICML*.
[63] Joseph Weizenbaum. 1972. On the Impact of the Computer on Society. *Science* 176, 4035 (1972), 609–614.
[64] Langdon Winner. 1977. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. MIT Press.
[65] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1 (1980), 121–136.