Unpacking the Expressed Consequences of AI Research in Broader Impact Statements

Priyanka Nanayakkara
priyankan@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Jessica Hullman
jhullman@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University
Evanston, Illinois, USA

ABSTRACT

The computer science research community and the broader public have become increasingly aware of negative consequences of algorithmic systems. In response, the top-tier Neural Information Processing Systems (NeurIPS) conference for machine learning and artificial intelligence research required that authors include a statement of broader impact to reflect on potential positive and negative consequences of their work. We present the results of a qualitative thematic analysis of a sample of statements written for the 2020 conference. The themes we identify broadly fall into categories related to how consequences are expressed (e.g., valence, specificity, uncertainty), areas of impacts expressed (e.g., bias, the environment, labor, privacy), and researchers' recommendations for mitigating negative consequences in the future. In light of our results, we offer perspectives on how the broader impact statement can be implemented in future iterations to better align with potential goals.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Social and professional topics → Codes of ethics**.

KEYWORDS

AI ethics, broader impacts, anticipatory governance, thematic analysis

1 INTRODUCTION

Scientists and the broader public have long grappled with the scientist's role in considering the societal consequences of their work. According to philosopher of science Heather Douglas [16], scientific thinking since the 1960s has tended to embrace the notion of a value-free ideal, limiting the extent to which scientists engage with non-epistemic social, ethical, or political values in the scientific process. Yet, however well-established, this value-free ideal has failed to address the many ways in which such values invariably infiltrate the scientific enterprise, including how a scientist might make changes to their research agenda based on potential societal consequences.

While the idea of values in design and technology is hardly new (e.g., [18, 43]), the broader computer science community has recently begun to make more concerted attempts to challenge the value-free ideal. In particular, computer scientists have called for researchers to consider the downstream consequences of their work as part of the peer-review process [26], formally integrating the act of reflecting on both positive and negative societal consequences into the scientific enterprise. Suggestions like this one are timely, as the computer science research community—as well as the broader public—is becoming increasingly aware of the ways in which deployed technologies have disproportionate negative impacts on marginalized communities [7, 17, 32] and pose significant costs to the environment [6, 39, 41]), and is demanding scrutiny to account for negative consequences [15].

In 2020, the Conference on Neural Information Processing Systems (NeurIPS), a top-tier conference for machine learning (ML) research, required that authors submit a broader impact statement as part of each paper submission. As per official guidance, the statement was meant to include both positive and negative potential societal consequences. NeurIPS's broader impact requirement mirrored the call put forth by Hecht et al. [26], and as a result of some ambiguity in its messaging has been termed by Hullman [27] and others as an "experiment," ostensibly in facilitating more active and intentional engagement within computer science around societal consequences of research and technology.

One way to conceive of the act of writing the broader impact statement is as an "ethical tool," as defined by Moula and Sandin [31]; that is, "a practical method and/or conceptual framework with the main purpose of helping the user(s) improve their ethical deliberations in order to reach an ethically informed judgement or decision." Similarly, the broader impact statement is relevant to the ongoing conversation around Algorithmic Impact Assessments (AIAs) [30], which refer to methods of increasing accountability around algorithmic systems. More broadly, such impact statements may contribute to frameworks of Responsible Research and Innovation (RRI) [33, 38] which help govern the R&D process in ways that are responsive to ethical and societal concerns.

While the intended outcomes of the broader impact statement, as envisioned by the NeurIPS conference organizers, are ambiguous, the Hecht et al. [26] proposal suggests goals such as increased transparency of impacts for the community, and encouragement of reflection and research on ways to mitigate negative impacts. In this paper we examine the content of NeurIPS broader impact statements in light of these goals. Do broader impact statements

capture a wide array of positive and negative consequences? Is there evidence that authors are considering ways to mitigate negative impacts? We present a qualitative thematic analysis of hundreds of NeurIPS 2020 broader impact statements, characterizing the impacts—both positive and negative—and recommendations for mitigating negative consequences that researchers discussed in their statements. Our underlying goal is to gain insight into what and how researchers elaborated in their broader impact statements, with an eye toward how our results may inform similar future initiatives.

2 BACKGROUND

The broader impact statements we study operate as a governance tool within the peer review process. Thus, we first briefly situate our work within ideas of responsible research before connecting to literature on values in science and technology.

2.1 Responsible Research

RRI can be understood as a process in which stakeholders "become mutually responsible to each other and anticipate research and innovation outcomes" [38]. This in turn relates to models of anticipatory governance in which emerging technologies are steered in order to be adapted to societal needs and ethical considerations [23]. In general, there are a range of approaches to RRI and anticipatory governance including scenario-based methods which can help systematically explore potential outcomes, and participatory methods which encourage the inclusion of societal stakeholders and feedback into the research process [12]. In this work we focus on impact assessments as a mode of anticipatory governance that allows for evaluation and engagement with the ethical implications of a given technology. For example, Wright [44] introduces a framework for ethical impact assessments which includes five guiding ethical principles/dimensions (e.g., Respect for Autonomy, Nonmaleficence, etc., which relate to earlier work by Beauchamp and Childress [5]) as well as an accompanying set of ethical tools and procedures to consider a technology as it relates to the guiding dimensions.

AIAs, as previously mentioned, are impact assessments specifically within an algorithmic context. Moss et al. [30] outline several remaining questions about the details of how AIAs work in practice. It is possible to imagine broader impact statements functioning as an ethical tool within a larger AIA framework. However, it is also important to note that the broader impact statement alone, as defined by NeurIPS, does not explicitly require wider engagement with the public or communities that are likely to experience harms that result from a certain technology's deployment or use. Metcalf et al. [29] point out that this could result in abstract discussions of impacts that do not reflect impacts that are likely to be realized, suggesting that the statement may work better as one potential approach among other ethical tools used to surface relevant ethical issues around a given technology.

2.2 Values in Technology

In addition to philosophical inquiry into the role of values in science [13, 16], there is a substantial body of work that lays the foundation for discussing the values embedded in technology. This includes work such as Winner [43]'s re-envisioning of common

artifacts as political ones, work by Friedman [18] and others on value-sensitive design, an approach to design that is closely intertwined with a consideration of values that may be built in to a technology, and Barbrook and Cameron [4]'s writing on the combination of values of the New Left and the New Right into a particular kind of "Californian Ideology" manifested in Silicon Valley's technological innovations.

There is also a growing body of work that specifically examines values within the computer science community, including how those values are ultimately reflected in new technologies. Rogaway [36] describes values that are implicit specifically within cryptographic work, and how these values may have shifted since cryptography's origins. Hanna and Park [24] question the value of scalability in computer science, while Birhane et al. [9] provide a study of prominent values found in the ML literature. In terms of ethical deliberation around technology, Shen et al. [40] propose "Value Cards," a toolkit intended to facilitate deliberation around technology and societal values. Within the specific domain of artificial intelligence (AI), recent attention has been given to value alignment challenges and whether values should be embedded or learned from data [37].

3 THE NEURIPS 2020 BROADER IMPACT STATEMENT

In this section, we offer context on the NeurIPS 2020 broader impact statement requirement as background to our study. Authors submitting papers to NeurIPS in 2020 were required to also submit a broader impact statement, which would not count toward the official page limit. The official requirement from the call for papers[1] is sparse, including the following instructions: *"In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes."*

While the Frequently Asked Questions (FAQ) page[2] explains that a submission could not "be rejected solely on the basis of the Broader Impact Statement," the call for papers states that "[s]ubmissions will also be considered on ethical grounds" and that "a submission may be rejected for ethical considerations" The conflict between these statements invites confusion around the role of the broader impact statement in determining acceptance status [27]. Beyond the broader impact statements, the conference included an additional ethical review process, through which peer reviewers could flag papers for ethical concerns which would lead to thorough ethical reviews of these papers by designated ethics reviewers [28]. While official communication regarding ethics-related rejections is limited, it is plausible to think that broader impact statements could have played a role in the paper flagging process and subsequent ethics review.

Official guidance for writing the broader impact statement was limited, but included a note on writing statements on theoretical work that the authors believe do not have foreseeable societal impact. In particular, the FAQ page states that authors can write in

their statement that "This work does not present any foreseeable so-cietal consequence." Additionally, the conference webpage provides links to the call for broader impact statements to be incorporated into the peer-review process by Hecht et al. [26], two unofficial guides for statement writing [3, 25], and a NeurIPS paper published in 2019 [20] which included a concluding statement very similar to a broader impact statement. While Hecht [25]'s guide provides high-level suggestions, such as to begin writing the broader impact statement early, Ashurst et al. [3]'s guide provides authors with guiding questions, an "Impact Stack" to help facilitate systematic reflection on the layers of impact of a given technology, and multiple example statements. More broadly, Prunkl et al. [35] discuss benefits, risks, and challenges of a NeurIPS-like broader impact statement, and describe how the NeurIPS broader impact statement fits within a larger context of ethics practices (e.g., IRBs).

The program chairs of 2020's conference provide preliminary details [28] on the outcomes of the new broader impact requirement. In particular, they write that "about 9% of the submission[s] did not have such a section, and most submissions had a section with about 100 words." It is important to note that even though some submissions omitted the broader impact section, all camera-ready accepted papers were required to include it. Abuhamad and Rheault [1] conducted a small preliminary survey ($N = 50$) of researchers, the majority of whom submitted to NeurIPS 2020, and found that the authors surveyed expressed "nonchalance" toward the broader impact statement and saw it as a "burden," although the majority also spent less than two hours on composing the statement itself. This research also found that there was confusion around how statements would be assessed, and that authors would have liked example statements as guidance. Additionally, Boyarskaya et al. [10] conducted an initial analysis ($N = 35$) of NeurIPS 2020 broader impact statements that were available in paper pre-prints before publication of the conference pre-proceedings. They identified potential limitations of the broader impact statements in order to inform recommendations for future methods of anticipating harms.

4 METHODOLOGY

In this section we describe the data and analytic approach to our qualitative analysis of broader impact statements.

4.1 Data

We first created a dataset including the 1,899 papers in the pre-proceedings of NeurIPS 2020. For each of these papers, we recorded the title, author name(s), and link to the full-length PDF by parsing https://papers.nips.cc/paper/2020. We then randomized the order of papers and extracted broader impact statements from the PDFs as needed for the analysis. We used regular expressions to extract all text after the heading for "Broader Impact" and before a heading for either "Acknowledgements" or "References"; this method typically extracted the text for the broader impact statement of a given paper. In cases where the automated extraction method failed (e.g., when our parser did not account for a specific variation in heading titles), we manually recorded the text of the corresponding broader impact

statement. These extracted texts formed the corpus[3] that we used for subsequent analysis.

4.2 Analysis

We qualitatively analyzed broader impact statements using a thematic analysis approach [11, 19]. We began by sequentially open coding excerpts from the randomly ranked set of collected impact statements. We allowed themes to emerge from the data and followed the method of constant comparison to continuously assess the fit and organization of excerpts into stable conceptual groupings that were triangulated across statements. To code, the first author read statements while pulling out excerpts that fit into existing thematic clusters of statements, or excerpts that could potentially be the basis for new themes. Throughout the process, the first author considered whether a particular theme could be broken into multiple themes, or whether themes could be re-organized to make more conceptual sense. The first author also wrote memos to articulate and clarify intermediate understandings of themes as well as relationships between themes.

We also practiced analyst triangulation [34] in which all co-authors periodically met to discuss the conceptual organization of themes and whether excerpts falling under particular themes fit under said themes, updating the overall organization accordingly. After coding 300 statements we felt themes were largely saturated and stable with little new information by each next statement [21]. At this point, we stopped coding additional statements due to diminishing returns. We then revisited all 300 statements and tabulated the prevalence of selected themes.

5 RESULTS

We find that authors primarily use the broader impact statement to discuss consequences of their work and to at times offer suggestions for how to mitigate potential negative consequences. Because these topics naturally arise in the statements, we present our findings as they relate to these two broad categories: Impacts and Author Recommendations.

5.1 Impacts

In this section, we focus on features of broader impact statements that relate to potential consequences of research. Thus, we begin by describing *how* authors present potential consequences (Expression of Impacts). Then we describe *what* main categories emerged in terms of the types of impacts around which clusters formed (Types of Impacts), and the relationship authors describe between theoretical work and foreseeable societal consequences (Relationship to Theoretical Work). Finally, we present *who* authors say will be impacted by their work (Who is Impacted?) and *when* authors suggest society might see the consequences of their work (Timeframe of Impacts). We provide counts for selected themes beginning in Types of Impacts.

5.1.1 Expression of Impacts. Authors express impacts along multiple axes. Specifically, we identify dimensions describing the *valence,*

---

[3]https://docs.google.com/spreadsheets/d/1S3YAsFeuc0Aex_2geA7nD-qAm7FR8Lcu4eoMmEjUPVk/edit?usp=sharing

orientation, *specificity*, *uncertainty*, and *composition* of stated consequences.

Valence. We find that statements range in the extent to which they describe positive or negative consequences. We later describe themes that arise around valence with respect to composition, including patterns in how authors structure statements with regard to positive/negative consequences (see *Composition*).

Orientation. We find that the areas of impact authors describe vary in terms of their orientation toward more society-facing outcomes (e.g., increased surveillance) versus more technically-oriented outcomes (e.g., increased robustness of a system). Technically-oriented consequences often have potential impacts on society (e.g., increased robustness of a system used in the context of self-driving cars could reduce the number of errors made, which keeps people safe), however authors do not always explicitly note such connections.

Specificity. Authors express consequences with a range of specificity. At the more specific end of the spectrum, authors say their work can be used in particular ways in specific contexts such as education (e.g., "this work will be integrated as part of the tools that we intend to teach in Ph.D. courses" [91][†]). Authors also point to specific areas of research that their work will impact, specific ways in which their work will contribute to a certain line of research, or a specific type of person who will be impacted. In at least a few cases, authors provide detailed, domain-specific case studies describing how their work may impact society (e.g., Coston et al. [58][†] describe their work in the context of parole decisions), or give examples of how the completed work may have already impacted society (e.g., the environmental impacts of having trained neural network models [73][†]).

In terms of very general consequences, authors sometimes name broad impact areas (e.g., "This work primarily has applications in the fields of psychophysics and crowdsourcing, and more generally, in learning from human responses." [94][†]) but don't always explain how the work will impact those areas. Finally, some authors express the idea that societal outcomes are dependent on the use case, or are application-specific, so societal consequences are difficult to discuss.

Uncertainty. Authors contend with the limitations of their methods, and the uncertainty inherent in models, pointing these out in statements in the vein of negative consequences. These consequences are more technically-oriented, with gestures toward associated societal impacts. Some authors point out that if assumptions are not met or adequately considered, the system could lead to wrong or harmful conclusions (e.g., Moreno et al. [88], Squires et al. [103][†]). Additionally, authors acknowledge uncertainty around models by pointing out model assumptions. Hao and Orlitsky [66][†] note that "[a] potential downside is that the theoretical guarantees of the associated algorithms rely on the assumption correctness."

Authors are also careful to point out the lack of guarantees provided by their algorithms, saying that "it is also important for the buyer to beware; . . . there may remain cases of interest when our approximations fail without warning" [65][†] or that "we cannot guarantee that every constraint will hold" [90][†] as well as other

general limitations, such as that "density-ratio based methods are not a panacea; it is entirely possible for the technique to introduce new biases when correcting for existing ones" [98][†].

Finally, authors describe how there is a possibility of errors downstream in future training or deployment, and that the real-world behavior of a given system is uncertain. For example, Kumar et al. [75][†] write that "it can be difficult for a practitioner to predict the behavior of an algorithm when it is deployed" and Hoang et al. [68][†] write that "[w]hile applications of [their] work to real data could result in ethical considerations, this is an indirect (and unpredictable) side-effect of [their] work." Some authors describe specific ways in which there could be downstream errors. Mate et al. [86][†] write that "if the worker makes an error when entering data . . . then the algorithm could make the wrong recommendation" and Bhatia et al. [48][†] write that if a designer chooses an inappropriate target, it "could lead to incorrect inferences and unexpected behavior in the real world."

Composition. The composition dimension reflects the overall conceptual structure and mixture of material in a statement in terms of valence, orientation, specificity, and uncertainty.

Some statements contain a mixture of valences, including both positive and negative consequences (though not necessarily balanced equally). We also find statements that lean heavily toward positive consequences (that is, they mostly describe positive consequences but mention negative consequences briefly). On one extreme, we see that some authors *only* include positive consequences, omitting a discussion of negative consequences entirely. While some simply don't make any mention of negative impacts, others explicitly state that there are no negative consequences (e.g., "Our work does not have any negative societal impacts" [107][†]; "no one is put at disadvantage from this research" [61][†]).

In addition, we find a theme around the statement of negative consequences where some authors write that a particular technology could be misused by a malicious actor, leading to undesirable societal outcomes. In this sense, authors engage with uncertainty around how a technology could be used in the future. Scialom et al. [100][†] provide an example of potential misuse:

> . . . malicious actors can use the same technology to build tools detrimental to society, e.g. for creation and propagation of misleading (fake) news . . . , impersonation, and deceit.

In addition, we find some statements structured such that they predominantly reflect positive consequences, but include a brief caveat alluding to potential negative consequences; the caveat often appears near the end of the statement. Zhou et al. [117][†] describe the positive impacts of their work, then conclude with the following caveat: "However, any reinforcement learning method runs the risk of being applied to military activities. Our work is no exception."

We also find that some statements do not contain a discussion of societal consequences at all. The omission of consequences is typically stated simply: "This work does not present any foreseeable societal consequence," mirroring the instruction provided in the call for papers. In these cases, no reason is provided for the omission of consequences. In other cases, authors provide reasons around why there may be limited consequences of their work. For example, some authors write that because their work is based on

existing work, their contributions do not significantly change what is technologically possible. Along these lines, Biggs et al. [49][†] write about potential negative impacts of their work, but limit the extent to which the work introduces new types of impacts:

> Furthermore, our method is an improvement of existing capabilities, but does not introduce a radical new capability in machine learning. Thus our contribution is unlikely to facilitate misuse of technology which is already available to anyone.

Finally, we find that authors' statements are sometimes oriented solely toward technical consequences, only describing their work in the context of how it contributes to the research space. In these cases, the stated technical contributions are usually framed as positive consequences.

5.1.2 Types of Impacts. Our analysis finds several categories of impacts that authors tend to discuss in broader impact statements. Attributes of statements that fall into these categories vary in terms of the aforementioned dimensions of expression (e.g., *Composition, Valence*, etc.). In particular, we point out that these categories range in terms of their orientation (see Expression of Impacts: *Orientation*) toward societal versus technical consequences. For each type of impact presented below, we note the category's predominant orientation as it appeared in statements.

Privacy. Impacts around privacy ($19.3\%, N = 58$) are largely oriented toward societal outcomes including protection of personal data and increased surveillance. In terms of positive impacts, authors describe how their work facilitates local computation (therefore reducing the extent to which personal data is shared) (e.g., "we help facilitate on-device deep learning, which could replace traditional cloud computation and foster the protection of privacy" [52][†]), contributes to methods that rely on decentralization (e.g., "One benefit of decentralized algorithms is that it does not need the central node to collect all users' information and every node only communicates with its trusted neighbors" [80][†]), and contributes to other privacy-focused research areas such as differential privacy (e.g., Chen et al. [55][†]) which aim to further protect an individual's information.

On the other hand, authors note how their work could be detrimental to maintaining privacy. They write that their work could facilitate the identification of individuals (e.g.,"potential use cases may target users based on their activity patterns" [97][†]), contribute to surveillance methods (e.g., "some common concerns on [Deep Neural Networks (DNNs)] such as privacy breaching and heavy surveillance can be worsened by DNN devices that are more available economically by using our proposed techniques" [77][†]), or even "may raise issues in neuroethics, especially regarding mental privacy" [73][†].

Labor. We find that authors write about societally-oriented impacts of their research in terms of influences on employment/work ($6.3\%, N = 19$) and impacts to productivity. For example, they acknowledge that automation of manually-done processes could displace jobs. Mate et al. [86][†] write the following on the jobs of community health workers (CHWs):

> … we also present results that highlight our algorithm's ability to plan among thousands of processes

at a time, far more than for which a human could independently plan. Just making this capability available could encourage the automation of applicable interventions via automated calls or texts, potentially displacing CHW jobs …

Chao et al. [52][†] write more generally that "unemployment may increase due to the increased automation enabled by the deep learning." On the other hand, some authors state that while their work contributes to automation, it is intended to increase productivity (e.g., "These techniques are not meant to replace highly skilled human workers, but to help improve their productivity at work" [67][†]). Some authors explicitly describe the tension between automation reducing jobs versus increasing productivity:

> Like many AI technologies, when used in automation, our technology can have a positive impact (increased productivity) and a negative impact (decreased demand) on labor markets. [106][†]

Additionally, authors write about how their work could improve or facilitate work processes such as collaboration, for instance "by helping bridge the gap between the ML and the scientific computing communities through allowing them to share tools and more easily interoperate" [89][†] or by helping to bring multiple disciplines together (e.g., by introducing concepts from one field (or subfield) into another [110, 118][†]). Authors also describe how their work could be useful for decision-making tasks. In particular, they write about decision-making in the contexts of medicine and policy, and how their work could contribute to fairer decisions. Along these lines, Hu et al. [71][†] write the following:

> By adopting our method, decision makers can build multiple decision models simultaneously just from one historical dataset and ensure that all decision models will be fair after the deployment …

Environment. We find that authors describe societally-oriented impacts to the environment ($10\%, N = 30$). For instance, authors describe how decreases in computational costs for a given task that their work accomplishes could "[save] lots of energy consumption" [57][†] and "has the potential to reduce the carbon footprint of building AI models" [59][†]. In terms of other positive consequences, they write about how their work could be used to contribute to climate change models:

> By employing Gaussian processes for data assimilation and building them into larger frameworks, this could facilitate more accurate climate models compared to current methods … [50][†]

On the other hand, authors also note that the amount of energy necessary to implement their methods could be harmful to the environment:

> … the carbon footprint of the proposed approach remains low compared to many deep approaches used nowadays in machine learning applications … [74][†]

Media. Some authors seem to be aware of the impact their work could have on the media ($6\%, N = 18$). They especially focus on impacts around generating or spreading misinformation, through synthetic media such as images or video (i.e., deepfakes) or other

means. In this way, impacts around media are quite societally-oriented. For example, Phuoc et al. [95][†] draw attention to how their work could be used in service of fake news: "Similar to existing image editing software, this enables the creation of image manipulations that could be used for ill-intended misinformation (*fake news*)."

Not all authors discussing misinformation, however, describe their work as aiding the spread of misinformation. At least one set of authors describes how their work could mitigate the harmful effects of misinformation, as it "may be used to detect harmful articles which may contain fake news, violent content, and fraudulent materials" [54][†].

Bias. When it comes to impacts around bias (including fairness, [24%, $N = 72$]), authors write both about impacts in ways that are both societally and technically-oriented. We note that definitions of bias vary across these two orientations, and even within an orientation, sometimes take on slightly different meanings. We first describe impacts around bias that is expressed as societally-oriented and maps to historical forms of discrimination. We then describe how authors sometimes refer to bias in more technically-oriented ways—in these cases, they do not connect bias of a system to societal discrimination.

Bias that can manifest in discriminatory algorithmic decisions is frequently mentioned in broader impact statements. In particular, authors engage with the idea of biased datasets, which may refer to training data populations in which some groups are underrepresented (see [42]), though few authors define what they mean by dataset bias in their statements. Some write that biased datasets are not an issue in their work; for instance, Zhang et al. [115][†] write that they "validate [their] method on large-scale public vision-language datasets and do not leverage biases in the data." Others describe how biases in datasets (stemming from society at large) may play a role in their work or the impacts of their work:

> . . . because the majority of [genome-wide association studies] are performed on individuals of European ancestries, [polygenic scores] are more accurate for individuals from those ancestry groups, potentially exacerbating health disparities between individuals of different ancestries . . . [102][†]

Additionally, authors write about discriminatory bias in the context of models. For instance, Liu and Scarlett [82][†] write that the signal recovery model of interest (which takes as input a separate model) could inherit potential biases of the input model. Zhang and Sabuncu [114][†] describe the impacts of bias potentially introduced during training:

> . . . our finding advocates for the use of priors for the regularization of neural networks. Despite the potentially better generalization performance of trained models, depending on the choice of priors used for training, unwanted bias can be inevitably introduced into the deep learning system, potentially causing issues of fairness . . .

On the other hand, authors talk about how their work could be used to mitigate these biases in algorithmic systems or datasets. For example, Chen et al. [56][†] describe how their work helps address dataset bias:

> . . . our theoretical work guides efforts to mitigate dataset bias. We demonstrate that curating a diverse pool of unlabeled data from the true population can help combating existing bias in labeled datasets. We give conditions for when bias will be mitigated and when it will be reinforced or amplified by popular algorithms . . .

At times, authors write about impacts around bias in a technical sense. For example:

> First, we note that finding clusters in networks is critical to *reducing bias* on measuring interventions with network effects . . . Having more flexible and better ways of doing this clustering will improve our ability to assess treatments on networks. [79][†]

It is, however, not always clear whether authors are referring to bias in a societal or technical sense, or whether technical forms of bias are related to societal inequities. To illustrate this point, we point to Xie et al. [109][†]'s description of how failure in their system could lead to biases in downstream systems:

> On the negative aspect, as many imaging-based research tasks and computer vision applications may be built upon the denoising algorithms, the failure of [the paper's framework for denoising] could potentially lead to biases or failures in these tasks and applications.

The authors do not specify the form of bias they have in mind, which underscores an aspect of the dimension of specificity in the impact statements. In cases such as this, terminology could be understood in different ways and has not been precisely articulated and disambiguated.

Efficiency. Efficiency (30%, $N = 90$) is a predominantly technically-oriented attribute that appears to often be prioritized in these research contexts, though justifications for it vary if given at all. However, authors do sometimes map the impact of increased/decreased efficiency to societal outcomes. For instance, authors write about how the increased efficiency afforded by their work (e.g., in terms of decreased training time) could save either time, money, or other resources. As an example, Zhou et al. [116][†] write about how their work could be applied to achieve "efficient package delivery with swarms of drones to reduce delivery costs." In addition, authors write about how increased efficiency demonstrated by their work could be beneficial for the environment, in terms of reduced computational costs, an idea that we see in the *Environment* theme as well. Some authors connect increased efficiency to the democratization of technology (see the *Democratization* theme below). For example, Larsson et al. [76][†] write that "the improved efficiency resulting from the predictor screening rules will make the [Sorted L-One Penalized Estimation (SLOPE)] family of models . . . accessible to a broader audience, enabling researchers and other parties to fit SLOPE models with improved efficiency" and go on to note that their work makes it possible to analyze "data sets that were otherwise too large to be analyzed without access to dedicated high-performance computing clusters."

While some authors imply that efficiency is a positive outcome, they may not explicitly draw a connection between increased efficiency and a particular societal outcome of efficiency. In these cases,

efficiency as an impact functions in very much of a technically-oriented capacity. For example, Brennan et al. [51][†] write about improving the effectiveness of a certain method, then write that they "hope to make [the method] more tractable, efficient, and broadly applicable"; because they don't make an explicit connection between efficiency itself and a particular outcome, it is difficult to make conclusions about the reasoning behind efficiency in particular. Finally, there are cases where authors invoke the role of efficiency in enabling other impacts such as those related to privacy or increased surveillance (e.g., "there are various unexpected yet feasible negative collateral consequences of increased efficiency, e.g., in terms of privacy" [46][†]).

Generalizability. Authors write about the generalizability of their methods (15.3%, $N = 46$) as positive consequences of their work, orienting generalizability technically, but sometimes connecting the technical contribution to a potential societal consequence. While it seems generally accepted that generalizability leads to positive outcomes (or is positive in and of itself), Hu et al. [70][†] highlight the tension between generalizability and potential application to societally undesirable use cases. In particular, they write that "[their] method has the potential to generalize the existing AI algorithms to more applications, while it also raises serious concern of privacy" going on to acknowledge the potential negative consequences that could arise from applying their work to the online behaviors of web users.

Authors also write about scalability as an important outcome or limitation of their work. For example, Yang et al. [110][†] note that while they improve scalability, their improvements are not quite enough, as it is still not possible to "scale to environments with thousands of agents and hundreds of types where the state-action space has thousands of dimensions."

Democratization. A theme emerges where researchers see their work as democratizing (5.7%, $N = 17$) by making some technology available to more people or organizations (e.g., in cases where previously using the technology required significant resources). Thus, we find that authors primarily characterize the increase in democratization of technology as a positive consequence of their work. In particular, they write about how their work enables more people in both research and non-research contexts to make use of various technologies. In this way, impacts around accessibility are both technically-oriented (e.g., improving researchers' accessibility to a particular technology enables further research in a given area) and societally-oriented (e.g., increasing accessibility of a particular technology enables wider use and participation by the wider public). More specifically, authors write about providing researchers with the ability to "harness volunteer computing and train models on the scale currently available only to large corporations" [99][†], while also discussing how their work can "help all the amateurs to use machine learning without any hassle" [113][†]. As more of an exception, Hong et al. [69][†] write directly about both the potential positive and negative consequences of increasing accessibility of technology:

> Techniques like [Spatially Stochastic Networks (SSNs)] that make content creation more accessible can help bridge this gap in representation [of developers] ... As any technology that promises to give easier access, it

has the potential for misuse. One could imagine cases where generative models are used to create things that may be harmful to society, and this can lower the technical entry barrier to misusers of this technology. For instance, SSNs could be applied toward harmful DeepFakes.

Robustness and Reliability. Authors describe impacts around robustness (e.g., against adversarial attacks) and reliability (21.3%, $N = 64$). In particular, they cite robustness as a positive attribute of their work, or lack of robustness as a limitation. While impacts in this area tend to be more technically-oriented, authors do draw connections to more society-facing impacts that stem from robustness. They note the importance of robustness in technologies that are used in contexts where safety is of concern, such as in "self-driving cars" [60][†]. Along the same lines of minimizing system errors, authors also write about their contributions with respect to increased reliability, including the ways in which their work can help people contend with uncertainty, noting that "for real-world classification tasks like disease diagnosis, in addition to accurate predictions, we need reliable estimates of the level of confidence of the predictions made" [114][†].

Accuracy. We find that the concept of accuracy (12.7%, $N = 38$) arises in statements. Authors write about improvements to accuracy as positive consequences of their work, mainly as technically-oriented consequences with allusions to societal consequences. For instance, they write that their work "could have significant broader impact by allowing users to more accurately solve practical problems" [68][†] or more simply stating that their "method ... can lead to even higher accuracies" which "is of high practical relevance" [85][†].

Interpretability. Authors describe impacts of their work that relate to interpretability (10.3%, $N = 31$). For instance, authors write about how their contributions to increased interpretability of algorithms or models can contribute to understanding biases or "promoting fairness in artificial intelligence" [112][†]:

> Moreover our algorithms have the additional benefit that they lead to more readily interpretable hypotheses ... This could potentially help practitioners better understand and diagnose complex machine learning systems they are designing, and troubleshoot ways that the algorithm might be amplifying biases in the data. [53][†]

Authors also write about how interpretability relates to the potential consequence of engendering trust in systems, which is framed as a positive consequence:

> In a positive prospect, we believe that our model contributes to further the development of less opaque machine learning models ... debugging and interpreting the model's behaviour and can help to establish trust toward the model when employed in larger application pipelines. [62][†]

In other cases, authors cite lack of interpretability as a negative consequence (or limitation) of their work. For example, Zhou et al. [116][†] say that "[their] method ... faces the 'black box problem' where behaviors of the individual agents may not be rational or interpretable from the human perspective." They go on to talk about

potential fairness issues, suggesting a connection between lack of interpretability and potential for societally negative outcomes.

5.1.3 Relationship to Theoretical Work. We find that authors can differ in how they frame the relationship between theoretical work and societal implications. A theme emerges where authors indicate, either explicitly or implicitly, that due to the theoretical nature of their work, there are *no* forseeable, negative, or ethical consequences (9%, $N = 27$). In particular, some narrow down the claim to be that there are no direct *negative* impacts of the work. However, this does not necessarily mean that they omit a discussion of other impacts of the work (positive or technical). Sometimes authors of theoretical work write that there are no foreseeable consequences altogether due to the nature of their contributions:

> The current paper presents theoretical work without any foreseeable societal consequence. Therefore, the authors believe that the broader impact discussion is not applicable. [63][†]

Note that these statements are slightly different from those that only state there are no foreseeable consequences (see Expression of Impacts: *Composition*), as they make a connection between theoretical work and lack of foreseeable consequences. We see another theme where authors write about the difficulty of foreseeing societal consequences of their work (e.g., "We believe that it is difficult to clearly foresee societal consequence of the present, purely theoretical, work" [83][†]).

Other authors mention the theoretical nature of their work, and write that there may be indirect consequences of their work or describe potential societal consequences (10%, $N = 30$). For example:

> Our results, while being theoretical in nature, have potential impacts in providing instructions for designing better security tools to ensure that people's online activities do not create unintended leakage of private information. [105][†]

5.1.4 Who is Impacted? In more than half of the broader impacts statements we analyzed, authors reference not only impacts, but also who might be impacted (64.3%, $N = 193$). These references vary in specificity. In a broad sense, authors often cited a domain (e.g., healthcare) which motivates their work or could be positively impacted by their work (e.g., Pinheiro et al. [96][†] write about the need for labeled data in medical imagery applications, and Yang et al. [111][†] write about how their work could help with "discovering the reasons for the quick spread of the epidemic disease in some areas"). At times, authors are more specific, mentioning particular groups who could be positively impacted, such as "people with language deficits like aphasia" [73][†] or people who could benefit from "[r]etinal prostheses" [84][†]. We also find a cluster of statements which mention impacts relating specifically to members of creative industries. For example, Tancik et al. [104][†] write that "this progress may also inadvertently reduce employment opportunities for the human artists who currently produce [photorealistic] effects manually," which fits into another theme of authors noting that those whose jobs may be automated could be negatively impacted (see Types of Impacts: *Labor*). Xie et al. [109][†] write about their

work on deep image denoising saying that "[i]ndividuals and corporations related to photography may benefit from [their] work," offering a potential positive consequence for a creative industry.

Other times, authors focus on impacts to members of various slices of the general public, such as those who use social media. For example, Zhu et al. [119][†] say that because "[their] work can be used in social media companies or any other occasions where user data can be accessed, people who are worried about their privacy being analyzed or targeted may be put at disadvantage." Authors also note how their work could positively or negatively impact people from historically disadvantaged groups, for example "by preventing them from receiving biased decisions" [71][†] or by being used "to monitor and negatively impact individuals in protected groups."

Lastly, authors turn their attention to researchers and practitioners. They write about how their work could benefit researchers both within computer science (e.g., "[o]ur study will ... benefit the machine learning community" [81][†]) and outside of computer science (e.g., atmospheric sciences [47][†]). We assume that when authors write about how their work could open up future research, they are implying that other researchers will be impacted. Authors also write that their work can help practitioners in their processes, such as by assisting them in "formally [reasoning] about the trade-offs across accuracy, robustness, and communication efficiency" [14][†].

5.1.5 Timeframe of Impacts. When authors write about the timeframe of impacts, that is, when society may see the consequences of the work at hand, they use broad terms that generally refer to a timeframe in terms of immediate versus long-term impact, and how their work might accelerate the realization of some outcome (10.3%, $N = 31$). For instance, they refer to work as having "no immediate practical applications" [87][†] or work that "can in the long run cause positive or negative societal impact" [106][†]. The level of detail in terms of timeframe typically does not go beyond immediate versus non-immediate results.

5.2 Author Recommendations

While authors tend to focus more attention on impacts, we also find that they sometimes make recommendations for how to mitigate potential negative consequences. We describe outcomes that authors suggest working toward, including their proposed steps for achieving these outcomes and their ascriptions about responsibility for taking action.

5.2.1 Outcomes. We describe five categories of desirable future outcomes that authors write about for the field and society to consider. In several cases, these outcomes align with positive impacts of their completed work, as described in Impacts: *Types of Impacts.*

Safe and Effective Use of AI. Authors make suggestions for how to ensure the safe or otherwise more effective use of their work, or AI in general (20.7%, $N = 62$). In particular, they suggest that "the practitioner verify how realistic these modeling assumptions are for the application at hand" [46][†] and note the importance of "[understanding] the principle and limitation of an algorithm to prevent failure" [92][†]. Authors also suggest more work around model robustness and write about the need for interpretability of models (e.g., with regards to algorithms used in safety-critical

environments [108][†]). Additionally, authors suggest human-in-the-loop methods (e.g., to check outputs in a personalized medicine context [72][†]), the need to develop methods of detecting misuses of technology (e.g., in the case of deepfakes [101][†]), and the potential need for regulation or policy around certain technologies (e.g., again in the case of deepfakes [69][†]).

Ensure "Fair" Outcomes. With regards to recommendations around issues of fairness (6.3%, $N = 19$), authors sometimes make general mentions of the importance of paying future attention to "fairness and nondiscrimination" [78][†] or the area of algorithmic fairness as it relates to their work [91][†]. Other times, authors make more specific suggestions, such as recommending to diversify tasks or datasets, saying, for instance, that "it is important to wisely choose the data on which the system is trained … If data is biased our method is not guaranteed to provide a correct estimation; this could harm the final users and should be carefully taken into account" [93][†].

Protect Privacy. Authors make recommendations to protect privacy (5.3%, $N = 16$). For example, authors make recommendations to "deploy [their] model locally" [101][†] and nod toward the importance of privacy-focused research areas such as differential privacy (e.g., Nguyen et al. [91][†]) moving forward.

Additionally, authors write more generally about the need for "privacy protections …throughout data collection, training, and deployment" [45][†], and broadly refer to legislative or regulatory means as ways of protecting privacy going forward. Foster et al. [64] write that "[they] welcome regulatory efforts to produce a legal framework that steers the usage of machine learning …in a direction which …respects …the privacy rights of users."

Reduce Environmental Impact. Similar to how authors bring up increased efficiency of their work as potentially benefiting the environment, they make recommendations for increased efficiency as a means of reducing environmental impact in the future (1%, $N = 3$).

5.2.2 Who is Responsible? We find that researchers sometimes make recommendations but do not assign responsibility (7.3%, $N = 22$), leaving it unclear who is ultimately responsible to take follow-up action. However, when authors do make mention of who might be responsible (either explicitly or implicitly) (24%, $N = 72$), they sometimes refer to various stakeholders, including people involved closer to deployment (e.g., practitioners, system designers, field experts, companies). They also refer to researchers (sometimes authors say that they are personally responsible for some action, like being involved in policy discussions around a certain technology; others leave their suggestions as a call to action for the broader research community; others describe recommendations for future computer science work implying that the research community is responsible) and policymakers.

6 DISCUSSION AND CONCLUSION

We find salient themes that emerge in broader impact statements from NeurIPS 2020; these findings indicate that in aggregate, authors discussed a fairly wide range of consequences including mentioning ideas for mitigating negative consequences. These descriptive findings contribute a set of dimensions and themes which could inform guidelines to shape what and how a research community

might frame a broader impact statement requirement so as to reduce ambiguity of intention that might have contributed to the wide range of themes we observed, such as by guiding expectations around authors' treatment of orientation, specificity, uncertainty, range of impacts, who is impacted, timeframe of impacts, and so on.

In closing, we elaborate a discussion based on how the statement appears to function in light of our results. We consider our findings through the lens of three potential goals: encouraging reflexivity, initiating changes to future research, and minimizing negligence and recklessness. Each offers opportunities for research communities such as NeurIPS to further steer or even evaluate broader impact statements.

Encouraging Reflexivity. One potential goal of the statement might have been simply to encourage researchers to reflect on the societal consequences of their work, or engage in moral inquiry, which Bietti [8] argues is intrinsically valuable. More formally, reflexivity as a political rationality found in the literature on social responsibility is "focused on making scientists aware of their own values and motivations, as well as making them reflect on the possible outcomes of their scientific inquiries" [22].

Our analysis finds evidence that the exercise surfaced values of the computer science research community, through the relatively frequent mentions of themes like efficiency, robustness, generalizability, or privacy, but it is less clear whether researchers devoted much effort to deliberating on their own personal values. In this sense, our findings contribute to the recent investigation into values of ML research by Birhane et al. [9], who identify common values present in influential ML publications such as accuracy, efficiency, and generalization. However, while they find that "overt consideration of societal benefits or harms is extremely rare," we find that in aggregate, broader impact statements do address societal consequences, suggesting that the statement requirement helps address an otherwise overlooked area. We did also observe some statements that reference technical contributions but do not necessarily map these contributions to potential *societal* impacts. If the goal is to encourage reflection on societal impacts, it could be useful for the official requirements/guidance around the statement to further encourage taking impacts an extra step further to consider how society could be impacted as a result of technical contributions (see Ashurst et al. [3]'s "Impact Stack") and to also encourage taking into account context when considering potential harms [10].

Initiate Changes to Future Research. Second, a potential goal of the broader impact statement might have been to help initiate changes in research agendas; the idea being that upon considering societal consequences of their work, researchers may have determined the potential negative consequences to far outweigh potential positive ones, prompting a shift in future research. Because significant time has not passed since authors wrote the statement, it is difficult to know whether it played any role in shifting research directions. It could be useful to observe whether content of statements shifts over time, and if discussions that arise in one year's set of statements manifest themselves in the work done in the following year(s). It could also be useful to conduct interviews with researchers to understand how, if at all, writing the statement informed consequent research questions.

Hecht [25] suggests that writing a broader impact statement early on in the research process leaves room to adapt a project accordingly, which is also in line with ideas from anticipatory governance [12]. Our analysis finds that researchers most likely completed their broader impact statements toward the end of the research process, namely because statements describe impacts of the work in more of a retroactive manner. In part, this could have been because the announcement of the broader impact requirement came at a time when work that was ultimately submitted was already later-stage. If the goal is to allow earlier-stage and continuous reflection, the requirements for the statement could benefit from clarification or perhaps changes in implementation to ensure researchers draft statements (or engage in ethical deliberation) early on in their research processes.

Minimize Negligence and Recklessness. Third, we explore the possibility that the broader impact statement may function as a way to minimize research negligence and recklessness. According to Douglas [16], "[r]ecklessness is proceeding in the face of unreasonable risk; negligence is the failure to foresee and mitigate such risk." The broader impact statement, therefore, can be seen as ensuring that authors engage with potential consequences, therefore reducing the chance of negligence. By making suggestions for mitigating negative consequences, researchers also minimize the risk of recklessness, in some way contributing to the mitigation of the risks of their own work by offering potential ways forward. While our findings already show a significant breadth of impacts, it could nonetheless be helpful to provide authors with particularly relevant ethical issues that relate to their general area of work, such as by referring them to established ethics tools (e.g., [44]), so that they may do a more complete job of assessing risks.

Finally, we note that several of the above perspectives imply that a practice like requiring broader impacts statements can be successful without drastic changes aimed at increasing the skills and knowledge of computer scientists around ethics and anticipation of societal outcomes of technology. For example, if the statement is intended to minimize negligence and recklessness, this would seem to imply that the community as a whole is largely capable of identifying the potentially harmful consequences of their work. Some have questioned, however, whether computer scientists are prepared to make useful observations about potential ethical problems with their work, and whether the results will be closer to "speculative fiction" [2]. Additionally, participatory ethics methods [12] suggest that community stakeholders likely to be affected should be involved in research to ensure that researchers are aware of possible problems they might not foresee. In the long run, it would seem that any expected instrumental values of broader impacts will depend on a clearly defined stance on the extent to which the community is intended to seek new skills, or collaborations, in reflecting on ethical implications.

ACKNOWLEDGMENTS

REFERENCES

[1] Grace Abuhamad and Claudel Rheault. 2020. Like a Researcher Stating Broader Impact For the Very First Time. *Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems* (2020).

[2] Eytan Adar. 2018. Negative Potential in CS. https://medium.com/@eytanadar/negative-potential-in-cs-b26964b4f502.

[3] Carolyn Ashurst, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. 2020. A Guide to Writing the NeurIPS Impact Statement. https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

[4] Richard Barbrook and Andy Cameron. 1996. The Californian Ideology. *Science as Culture* 6, 1 (1996), 44–72.

[5] Tom L Beauchamp and James F Childress. 2001. *Principles of Biomedical Ethics.* Oxford University Press, USA.

[6] EM Bender, T Gebru, A McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2021).

[7] Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces* (2019).

[8] Elettra Bietti. 2020. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 210–219.

[9] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2020. The Underlying Values of Machine Learning Research. *Resistance AI Workshop at NeurIPS 2020* (2020).

[10] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. *Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems* (2020).

[11] Virginia Braun and Victoria Clarke. 2008. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (July 2008), 77–101.

[12] Philip Brey. 2017. Ethics of Emerging Technology. Rowman and Littlefield International. https://ethicsandtechnology.eu/wp-content/uploads/downloadable-content/Brey-2017-Ethics-Emerging-Tech.pdf

[13] Matthew Brown. 2020. *Science and Moral Imagination: A New Ideal for Values in Science.* University of Pittsburgh Press.

[14] Xiangyi Chen, Tiancong Chen, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. 2020. Distributed Training with Heterogeneous Data: Bridging Median-and Mean-based Algorithms. *Advances in Neural Information Processing Systems* 33 (2020).

[15] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Jan. 2016), 56–62. https://doi.org/10.1145/2844110

[16] Heather Douglas. 2009. *Science, Policy, and the Value-Free Ideal.* University of Pittsburgh Press.

[17] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[18] Batya Friedman. 1996. Value-Sensitive Design. *Interactions* 3, 6 (1996), 16–23.

[19] Graham R Gibbs. 2007. *Analyzing qualitative data.* Sage London, England.

[20] Jon Gillick and David Bamman. 2019. Breaking Speech Recognizers to Imagine Lyrics. *Advances in Neural Information Processing Systems* 32 (2019).

[21] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research.* Routledge.

[22] Cecilie Glerup and Maja Horst. 2014. Mapping 'social responsibility' in science. *Journal of Responsible Innovation* 1, 1 (2014), 31–50.

[23] David H Guston. 2014. Understanding 'anticipatory governance'. *Social studies of science* 44, 2 (2014), 218–242.

[24] Alex Hanna and Tina M. Park. 2020. Against Scale: Provocations and Resistances to Scale Thinking. *Resistance AI Workshop at NeurIPS 2020* (2020).

[25] Brent Hecht. 2020. Suggestions for Writing NeurIPS 2020 Broader Impacts Statements. https://brenthecht.medium.com/suggestions-for-writing-neurips-2020-broader-impacts-statements-121da1b765bf

[26] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. 2018. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. *ACM Future of Computing Blog* (2018).

[27] Jessica Hullman. 2020. The NeurIPS 2020 broader impacts experiment. https://statmodeling.stat.columbia.edu/2020/12/21/the-neurips-2020-broader-impacts-experiment/.

[28] Hsuan-Tien Lin, Maria-Florina Balcan, Raia Hadsell, and Marc'Aurelio Ranzato. 2020. What we learned from NeurIPS 2020 reviewing process. https://neuripsconf.medium.com/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f

[29] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2021).

[30] Emanuel Moss, Elizabeth Anne Watkins, Jacob Metcalf, and Madeleine Clare Elish. 2020. Governing with Algorithmic Impact Assessments: Six Observations.

Available at SSRN (2020).

[31] Payam Moula and Per Sandin. 2017. Ethical Tools. In *The Ethics of Technology: Methods and Approaches*, Sven Ove Hansson (Ed.). Rowman & Littlefield, Chapter 8.

[32] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[33] R Owen, P Macnaghten, and J Stilgoe. 2012. Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy* 39, 6 (2012), 751–760. https://doi.org/10.1093/scipol/scs093

[34] Michael Quinn Patton. 2014. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (4 ed.). Sage Publications.

[35] Carina EA Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence* 3, 2 (2021), 104–110.

[36] Phillip Rogaway. 2015. The Moral Character of Cryptographic Work. *IACR Cryptol. ePrint Arch.* 2015 (2015), 1162.

[37] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[38] René von Schomberg. 2013. Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society. (2013), 51–74. https://doi.org/10.1002/9781118551424.ch3

[39] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (2020), 54–63.

[40] Hong Shen, Wesley Deng, Aditi Chattopadhyay, Steven Wu, Xu Wang, and Haiyi Zhu. 2020. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. *arXiv preprint arXiv:2010.11411* (2020).

[41] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[42] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002* (2019).

[43] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* (1980), 121–136.

[44] David Wright. 2011. A framework for the ethical impact assessment of information technology. *Ethics and Information Technology* 13, 3 (2011), 199–226.

APPENDIX: REFERENCES FOR BROADER IMPACT STATEMENTS

[45] Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems* 33 (2020).

[46] David Alvarez-Melis and Nicolò Fusi. 2020. Geometric Dataset Distances via Optimal Transport. *Advances in Neural Information Processing Systems* 33 (2020).

[47] Julius Berner, Markus Dablander, and Philipp Grohs. 2020. Numerically Solving Parametric Families of High-Dimensional Kolmogorov Partial Differential Equations via Deep Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[48] Kush Bhatia, Ashwin Pananjady, Peter L Bartlett, Anca Dragan, and Martin J Wainwright. 2020. Preference learning along multiple criteria: A game theoretic perspective. *Advances in Neural Information Processing Systems* 33 (2020).

[49] Benjamin Biggs, Sébastien Ehrhadt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 2020. 3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data. *Advances in Neural Information Processing Systems* 33 (2020).

[50] Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. 2020. Matérn Gaussian processes on Riemannian manifolds. *Advances in Neural Information Processing Systems* 33 (2020).

[51] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. 2020. Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems* 33 (2020).

[52] Shih-Kang Chao, Zhanyu Wang, Yue Xing, and Guang Cheng. 2020. Directional Pruning of Deep Neural Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[53] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2020. Classification Under Misspecification: Halfspaces, Generalized Linear Models, and Evolvability. *Advances in Neural Information Processing Systems* 33 (2020).

[54] Wenchao Chen, Chaojie Wang, Bo Chen, Yicheng Liu, Hao Zhang, and Mingyuan Zhou. 2020. Bidirectional Convolutional Poisson Gamma Dynamical Systems. *Advances in Neural Information Processing Systems* 33 (2020).

[55] Xiangyi Chen, Steven Z Wu, and Mingyi Hong. 2020. Understanding Gradient Clipping in Private SGD: A Geometric Perspective. *Advances in Neural Information Processing Systems* 33 (2020).

[56] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. 2020. Self-training Avoids Using Spurious Features Under Domain Shift. *Advances in Neural Information Processing Systems* 33 (2020).

[57] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. 2020. Hierarchical Neural Architecture Search for Deep Stereo Matching. *Advances in Neural Information Processing Systems* 33 (2020).

[58] Amanda Coston, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual Predictions under Runtime Confounding. *Advances in Neural Information Processing Systems* 33 (2020).

[59] Talgat Daulbaev, Alexandr Katrutsa, Larisa Markeeva, Julia Gusak, Andrzej Cichocki, and Ivan Oseledets. 2020. Interpolation Technique to Speed Up Gradients Propagation in Neural ODEs. *Advances in Neural Information Processing Systems* 33 (2020).

[60] Ilias Diakonikolas, Daniel M Kane, and Pasin Manurangsi. 2020. The Complexity of Adversarially Robust Proper Learning of Halfspaces with Agnostic Noise. *Advances in Neural Information Processing Systems* 33 (2020).

[61] Vu Dinh and Lam Si Tung Ho. 2020. Consistent Feature Selection for Analytic Deep Neural Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[62] Sébastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. 2020. RELATE: Physically Plausible Multi-Object Scene Synthesis Using Structured Latent Spaces. *Advances in Neural Information Processing Systems* 33 (2020).

[63] Yuval Emek, Ron Lavi, Rad Niazadeh, and Yangguang Shi. 2020. Stateful posted pricing with vanishing regret via dynamic deterministic markov decision processes. *Advances in Neural Information Processing Systems* 33 (2020).

[64] Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. 2020. Adapting to Misspecification in Contextual Bandits. *Advances in Neural Information Processing Systems* 33 (2020).

[65] Soumya Ghosh, William T Stephenson, Tin D Nguyen, Sameer K Deshpande, and Tamara Broderick. 2020. Approximate Cross-Validation for Structured Models. *Advances in Neural Information Processing Systems* 33 (2020).

[66] Yi Hao and Alon Orlitsky. 2020. Profile Entropy: A Fundamental Measure for the Learnability and Compressibility of Discrete Distributions. *Advances in Neural Information Processing Systems* 33 (2020).

[67] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. 2020. Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction. *Advances in Neural Information Processing Systems* 33 (2020).

[68] Quang Minh Hoang, Trong Nghia Hoang, Hai Pham, and David P Woodruff. 2020. Revisiting the Sample Complexity of Sparse Spectrum Approximation of Gaussian Processes. *Advances in Neural Information Processing Systems* 33 (2020).

[69] Sarah Jane Hong, Martin Arjovsky, Ian Thompson, and Darryl Barnhardt. 2020. Low Distortion Block-Resampling with Spatially Stochastic Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[70] Hengtong Hu, Lingxi Xie, Zewei Du, Richang Hong, and Qi Tian. 2020. One-bit Supervision for Image Classification. *Advances in Neural Information Processing Systems* 33 (2020).

[71] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fair Multiple Decision Making Through Soft Interventions. *Advances in Neural Information Processing Systems* 33 (2020).

[72] Xiangru Huang, Haitao Yang, Etienne Vouga, and Qixing Huang. 2020. Dense Correspondences between Human Bodies via Learning Transformation Synchronization on Graphs. *Advances in Neural Information Processing Systems* 33 (2020).

[73] Shailee Jain, Vy A Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander G Huth. 2020. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *Advances in Neural Information Processing Systems* 33 (2020).

[74] Noémie Jaquier and Leonel Rozo. 2020. High-dimensional Bayesian Optimization via Nested Riemannian Manifolds. *Advances in Neural Information Processing Systems* 33 (2020).

[75] Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. 2020. One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL. *Advances in Neural Information Processing Systems* 33 (2020).

[76] Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. 2020. The Strong Screening Rule for SLOPE. *Advances in Neural Information Processing Systems* 33 (2020).

[77] Dongsoo Lee, Se Jung Kwon, Byeongwook Kim, Yongkweon Jeon, Baeseong Park, and Jeongin Yun. 2020. FleXOR: Trainable Fractional Quantization. *Advances in Neural Information Processing Systems* 33 (2020).

[78] Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Zhaoran Wang, and Mladen Kolar. 2020. Provably efficient neural estimation of structural equation model: An adversarial approach. *Advances in Neural Information Processing Systems* 33 (2020).

[79] Meng Liu and David F Gleich. 2020. Strongly local p-norm-cut algorithms for semi-supervised learning and local graph clustering. *Advances in Neural Information Processing Systems* 33 (2020).

[80] Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. 2020. A Decentralized Parallel Algorithm for Training Generative Adversarial Nets. *Advances in Neural Information Processing Systems*

33 (2020).

[81] Yanli Liu, Yuan Gao, and Wotao Yin. 2020. An Improved Analysis of Stochastic Gradient Descent with Momentum. *Advances in Neural Information Processing Systems* 33 (2020).

[82] Zhaoqiang Liu and Jonathan Scarlett. 2020. The Generalized Lasso with Nonlinear Observations and Generative Priors. *Advances in Neural Information Processing Systems* 33 (2020).

[83] Clément Luneau, Jean Barbier, and Nicolas Macris. 2020. Information theoretic limits of learning a sparse rule. *Advances in Neural Information Processing Systems* 33 (2020).

[84] Gabriel Mahuas, Giulio Isacchini, Olivier Marre, Ulisse Ferrari, and Thierry Mora. 2020. A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons. *Advances in Neural Information Processing Systems* (2020).

[85] Andrés R Masegosa, Stephan S Lorenzen, Christian Igel, and Yevgeny Seldin. 2020. Second order PAC-Bayesian bounds for the weighted majority vote. *Advances in Neural Information Processing Systems* 33 (2020).

[86] Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Interventions. *Advances in Neural Information Processing Systems* 33 (2020).

[87] Alexander Meulemans, Francesco S Carzaniga, Johan AK Suykens, João Sacramento, and Benjamin F Grewe. 2020. A Theoretical Framework for Target Propagation. *Advances in Neural Information Processing Systems* 33 (2020).

[88] Alexander Moreno, Zhenke Wu, Jamie Yap, David Wetter, Cho Lam, Inbal Nahum-Shani, Walter Dempsey, and James M Rehg. 2020. A Robust Functional EM Algorithm for Incomplete Panel Count Data. *Advances in Neural Information Processing Systems* 33 (2020).

[89] William S Moses and Valentin Churavy. 2020. Instead of Rewriting Foreign Code for Machine Learning, Automatically Synthesize Fast Gradients. *Advances in Neural Information Processing Systems* 33 (2020).

[90] Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, and Wenshuo Guo. 2020. Approximate Heavily-Constrained Learning with Lagrange Multiplier Models. *Advances in Neural Information Processing Systems* 33 (2020).

[91] Viet Anh Nguyen, Xuhui Zhang, Jose Blanchet, and Angelos Georghiou. 2020. Distributionally Robust Parametric Maximum Likelihood Estimation. *Advances in Neural Information Processing Systems* 33 (2020).

[92] Seongmin Ok. 2020. A Graph Similarity for Deep Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[93] Massimiliano Patacchiola, Jack Turner, Elliot J Crowley, Michael O'Boyle, and Amos Storkey. 2020. Bayesian Meta-Learning for the Few-Shot Setting via Deep Kernels. *Advances in Neural Information Processing Systems* 33 (2020).

[94] Michaël Perrot, Pascal Mattia Esser, and Debarghya Ghoshdastidar. 2020. Near-Optimal Comparison Based Clustering. *Advances in Neural Information Processing Systems* 33 (2020).

[95] Thu Nguyen Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy J Mitra. 2020. BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images. In *Advances in Neural Information Processing Systems*, Vol. 33.

[96] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmaleck, Florian Golemo, and Aaron Courville. 2020. Unsupervised Learning of Dense Visual Representations. *Advances in Neural Information Processing Systems* 33 (2020).

[97] Sirisha Rambhatla, Xingguo Li, and Jarvis Haupt. 2020. Provable Online CP/PARAFAC Decomposition of a Structured Tensor via Dictionary Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[98] Benjamin Rhodes, Kai Xu, and Michael U Gutmann. 2020. Telescoping Density-Ratio Estimation. *Advances in Neural Information Processing Systems* 33 (2020).

[99] Max Ryabinin and Anton Gusev. 2020. Towards Crowdsourced Training of Large Neural Networks using Decentralized Mixture-of-Experts. *Advances in Neural Information Processing Systems* 33 (2020).

[100] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. ColdGANs: Taming Language GANs with Cautious Sampling Strategies. *Advances in Neural Information Processing Systems* 33 (2020).

[101] Siyuan Shan, Yang Li, and Junier B Oliva. 2020. Meta-Neighborhoods. *Advances in Neural Information Processing Systems* 33 (2020).

[102] Jeffrey P Spence. 2020. Flexible mean field variational inference using mixtures of non-overlapping exponential families. *Advances in Neural Information Processing Systems* 33 (2020).

[103] Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. 2020. Active Structure Learning of Causal DAGs via Directed Clique Tree. *Advances in Neural Information Processing Systems* 33 (2020).

[104] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *Advances in Neural Information Processing Systems* 33 (2020).

[105] Wei Tang, Chien-Ju Ho, and Yang Liu. 2020. Optimal Query Complexity of Secure Stochastic Convex Optimization. *Advances in Neural Information Processing Systems* 33 (2020).

[106] Elise van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. 2020. MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[107] Lemeng Wu, Bo Liu, Peter Stone, and Qiang Liu. 2020. Firefly Neural Architecture Descent: a General Approach for Growing Neural Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[108] Fan Xie, Alexander Chowdhury, M Kaluza, Linfeng Zhao, Lawson LS Wong, and Rose Yu. 2020. Deep Imitation Learning for Bimanual Robotic Manipulation. *Advances in Neural Information Processing Systems* 33 (2020).

[109] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. 2020. Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising. *Advances in Neural Information Processing Systems* 33 (2020).

[110] Fan Yang, Alina Vereshchaka, Changyou Chen, and Wen Dong. 2020. Bayesian Multi-type Mean Field Multi-agent Imitation Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[111] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. 2020. Factorizable Graph Convolutional Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[112] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional Few-Shot Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[113] Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, and Steven Su. 2020. Differentiable Neural Architecture Search in Equivalent Space with Exploration Enhancement. *Advances in Neural Information Processing Systems* 33 (2020).

[114] Zhilu Zhang and Mert R Sabuncu. 2020. Self-Distillation as Instance-Specific Label Smoothing. *Advances in Neural Information Processing Systems* 33 (2020).

[115] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. *Advances in Neural Information Processing Systems* 33 (2020).

[116] Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. 2020. Learning Implicit Credit Assignment for Cooperative Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[117] Qi Zhou, Yufei Kuang, Zherui Qiu, Houqiang Li, and Jie Wang. 2020. Promoting Stochasticity for Expressive Policies via a Simple and Efficient Regularization Method. *Advances in Neural Information Processing Systems* 33 (2020).

[118] Yufan Zhou, Changyou Chen, and Jinhui Xu. 2020. Learning Manifold Implicitly via Explicit Heat-Kernel Learning. *Advances in Neural Information Processing Systems* 33 (2020).

[119] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. 2020. Fewer is More: A Deep Graph Metric Learning Perspective Using Fewer Proxies. *Advances in Neural Information Processing Systems* 33 (2020).