Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews

David Liu[*]
Northeastern University
Boston, MA, USA
liu.davi@northeastern.edu

Priyanka Nanayakkara[*][†]
Northwestern University
Evanston, IL, USA
priyankan@u.northwestern.edu

Sarah Ariyan Sakha
International Rescue Committee
New York, NY, USA
s.sakha@columbia.edu

Grace Abuhamad
ServiceNow
Montréal, QC, Canada
abuhamad@alum.mit.edu

Su Lin Blodgett
Microsoft Research
Montréal, QC, Canada
sulin.blodgett@microsoft.com

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

Jessica R. Hullman
Northwestern Univeristy
Evanston, IL, USA
jhullman@northwestern.edu

Tina Eliassi-Rad
Northeastern University
Boston, MA, USA
tina@eliassi.org

ABSTRACT

The artificial intelligence research community is continuing to grapple with the ethics of its work by encouraging researchers to discuss potential positive and negative consequences. Neural Information Processing Systems (NeurIPS), a top-tier conference for machine learning and artificial intelligence research, first required a statement of broader impact in 2020. In 2021, NeurIPS updated their call for papers such that 1) the impact statement focused on *negative* societal impacts and was not required but encouraged, 2) a paper checklist and ethics guidelines were provided to authors, and 3) papers underwent ethics reviews and could be rejected on ethical grounds. In light of these changes, we contribute a qualitative analysis of 231 impact statements and all publicly-available ethics reviews. We describe themes arising around the ways in which authors express agency (or lack thereof) in identifying or mitigating negative consequences and assign responsibility for mitigating negative societal impacts. We also characterize ethics reviews in terms of the types of issues raised by ethics reviewers (falling into categories of policy-oriented and non-policy-oriented), recommendations ethics reviewers make to authors (e.g., in terms of adding or removing content), and interaction between authors, ethics reviewers, and original reviewers (e.g., consistency between issues flagged by original reviewers and those discussed by ethics reviewers). Finally, based on our analysis we make recommendations for how authors can be further supported in engaging with the ethical implications of their work.

CCS CONCEPTS

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

KEYWORDS

AI ethics, impact statements, broader impact, ethics review

ACM Reference Format:
David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R. Hullman, and Tina Eliassi-Rad. 2022. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534155

1 INTRODUCTION

Over the past few years, the artificial intelligence research community has been experimenting with the development of professional ethics norms, specifically around standards for reflecting on the societal implications of research prior to publication. In this vein, in 2018, Hecht et al. [15] proposed a change to the peer review process: researchers should be expected to include some discussion of both positive and negative societal consequences of their work in their submissions. That year, there was a workshop at a top-tier machine learning conference, Neural Information Processing Systems (NeurIPS), on ethical, social, and governance issues in artificial intelligence [4], and similar ones each year since [2, 12, 14, 25]. In 2020, NeurIPS announced more interdisciplinary subject areas [13] and a new submission requirement: a broader impact statement [18]. Unlike workshops and interdisciplinary tracks that might have been considered more "niche," the requirement directly affected every submission, of which there were over 9,000 in 2020 [1, 17]. While broader impact statements themselves were not new to the research community at-large [9, 22, 26], they were new to the NeurIPS community.

Multiple papers sought to analyze the response and effect of the 2020 NeurIPS submission requirement [1, 3, 10, 19], particularly

[*] Both authors contributed equally to this research.
[†] The author completed part of this work while at Columbia University as a visiting researcher.

AIES'22, August 1–3, 2022, Oxford, United Kingdom
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534155

to inform considerations for designing its next iteration. In 2021, the NeurIPS submission requirements changed: a broader impact statement was no longer required, but a "paper checklist" was [8, 20, 21]. The checklist included, and went beyond, the broader impact statement "to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact" [21]. Both the conference-provided ethics guidelines and paper checklist included examples of potential negative societal impacts and provided researchers with suggestions for considerations of impacted stakeholders and possible harms [5, 21]. The checklist also encouraged authors to discuss mitigation strategies where there was potential negative societal impact [21]. In addition to the paper checklist, NeurIPS empowered its ethics review process—while a paper could not be rejected solely on ethical grounds in 2020, this became a possibility in 2021 [8]. Overall, NeurIPS 2021 program chairs viewed the ethics review process as "educational, not prohibitive" [7], and provided the paper checklist, ethics guidelines, and ethics review process as tools for ethical reflection.

Standards for how the artificial intelligence research community considers downstream impacts of their work are evolving and require closer examination. While previous work has, for example, studied the topics researchers discuss when required to write broader impact statements [19], changes to the impact statement requirement motivate further investigation into how authors approached the task. Further, the new availability of ethics reviews provides an opportunity to study the dialog between authors and ethics reviewers.

In this paper, we analyze the 2021 potential negative impact statements ("impact statements") and ethics reviews with the goal of understanding to what extent, and how, authors discuss negative impact when not explicitly required to do so, and the role of ethics reviews in encouraging authors to further consider the impacts of their work. We contribute a qualitative analysis of 231 impact statements, a 30% random sample of all statements collected, and all available ethics reviews (96 reviews across 50 papers). We describe themes arising around the ways in which authors express agency (or lack thereof) in identifying or mitigating negative societal impacts and assign responsibility for mitigating negative impacts in impact statements. These patterns provide further evidence for previously mentioned barriers to accountability, such as the difficulty of placing blame when "many hands" are involved in the creation of machine learning systems and the characterization of bugs as inevitable [11]. We also describe themes in the ethics review process centering around the types of issues raised by ethics reviewers, recommendations ethics reviewers make to authors, and interaction between authors, ethics reviewers, and original reviewers.

2 DATA AND METHODS

Our process for analyzing both the impact statements and the ethics reviews followed three similar stages 1) data collection, 2) taxonomy specification, and 3) qualitative coding, which are further broken down in Figure 1. We discuss our data collection methods, highlighting our impact statement scraping process, and provide details on our qualitative coding methodologies for impact statements and ethics reviews.

2.1 Impact Statement Scraping

Two changes to 2021's impact statement policy informed our data collection process. First, papers were not required to include an impact statement; instead NeurIPS encouraged authors to discuss potential negative societal impacts, but omission of a statement was not grounds for rejection [5]. Second, discussions of societal impacts did not need to have their own dedicated section. We designed our statement scraping pipeline, shown in Figure 2, to accommodate these changes.

First, we downloaded the camera-ready paper PDFs from the NeurIPS Pre-Proceedings.[1] We converted these PDFs into HTML using `pdfminer`[2] and parsed the output with `BeautifulSoup`.[3]

To extract the impact statements, we defined start and end elements in the HTML files. The start element is the first element containing any of the following key phrases: "broader impact," "societal impact," "broad impact," "social impact," and "negative impact." We devised the list of key phrases based on the initial manual validation process detailed in Section 2.3. Papers without a start element were deemed to not contain an impact statement. The end element is the first element after the start that contains a section header, as determined by the font (since header fonts differ from other parts of paper text); while any section header ends the statement,[4] we explicitly checked for the beginning of a references or acknowledgements section since impact statements often appear near the end of papers. The final statement we extract is then the concatenation of all text between the start and end elements. We also repeat the process for the supplemental materials to extract impact statements appearing in the appendix, though we only process supplemental materials that are PDFs and omit zip folders.

2.1.1 *Scraper Validation.*

First, we manually validated the initial scraper by comparing extracted statements against the original PDFs for 72 papers. We further validated our scraper by checking for false negatives, which are discussions of societal impact that do not contain any of the key phrases listed in Figure 2. To account for false negatives, we took a random sample of a dozen papers from among the 1, 566 our scraper did not select. We read each of these papers checking for any discussion of societal impact and found no false negatives. The closest discussions we found used real-world applications as motivation in the introduction but did not go further to discuss societal impact.

2.1.2 *Data Summary.*

In total, we found impact statements in 768 papers or 32.9% of papers. Among the papers that did discuss societal impact, 80.7% included a statement in the main paper body and 28.9% included a statement in the appendix, with some including statements in both. Among papers that had impact statements, the median length was similar to the length of 2020 NeurIPS statements [3]. The median length for statements included in the main body was 113 words while the median length was 136 in 2020.

[1] https://proceedings.neurips.cc/paper/2021 Accessed on November 15, 2021
[2] https://github.com/pdfminer/pdfminer.six
[3] https://pypi.org/project/beautifulsoup4/
[4] If the statement begins with a section header, we ensure the end header is not a child header to prevent trimming the statement.

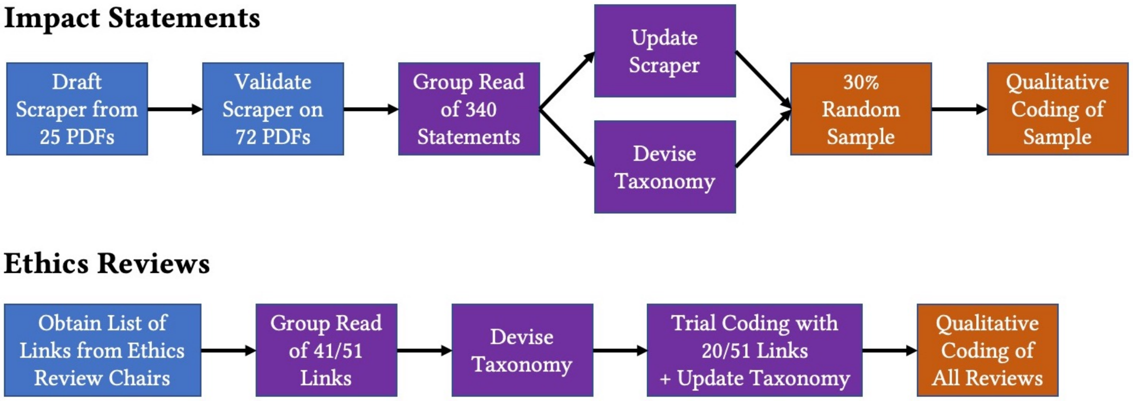
Impact Statements
Draft Scraper from 25 PDFs
Validate Scraper on 72 PDFs
Group Read of 340 Statements
Update Scraper
Devise Taxonomy
30% Random Sample
Qualitative Coding of Sample
Ethics Reviews
Obtain List of Links from Ethics Review Chairs
Group Read of 41/51 Links
Devise Taxonomy
Trial Coding with 20/51 Links + Update Taxonomy
Qualitative Coding of All Reviews

Figure 1: We qualitatively coded both the impact statements and ethics reviews. Our processes for both sets of documents was broken into three stages: 1) data collection (blue) 2) taxonomy specification (purple) 3) final coding (orange). In both cases, we started the taxonomy specification process by reading a sample of documents as a group; we then discussed trends that we individually noticed and defined recurring themes. For the impact statements, we used the group read to identify lingering issues with the scraper and for the ethics reviews, we conducted a coding trial with our draft taxonomy given the small dataset size. Finally, for impact statements we coded a 30% random sample (231 statements) of all statements whereas we read all 96 ethics reviews.

2.2 Ethics Reviews

Because NeurIPS used the OpenReview platform for paper reviews in 2021, we were able to access all of the ethics reviews for papers that were accepted along with rejected papers that opted to make their reviews public. We received a list of 51 OpenReview links[5] that were publicly flagged for ethics concerns from the NeurIPS ethics chairs which yielded 96 ethics reviews. All accepted papers that underwent an ethics review are included in our dataset; however, there were 215 rejected papers with ethics reviews that are not public and thus are not in our dataset [6].

To contextualize the ethics review corpus, we summarize the 2021 NeurIPS ethics review process in Figure 3. For every submission, each NeurIPS reviewer, who we refer to as an "original" reviewer to distinguish from an ethics reviewer, had the ability to flag the paper as needing an ethics review. The original reviewers could choose among a set of pre-defined ethics categories and elaborate on their reasoning for flagging the paper. Thereafter, ethics reviewers would join the reviewing discussion and comment more extensively on ethical issues, if any, in the paper. It is important to note that by the time ethics reviews were added, authors could already read the original reviews and scores. Authors were given the chance to respond to the ethics reviews, and after the discussion period the area chair holistically assessed both original and ethics reviews to make final paper decisions. In rare cases, area chairs could make a "conditional accept" decision which would condition acceptance on the authors incorporating comments from the ethics review.

2.3 Qualitative Coding

For both the impact statements and ethics reviews, we followed an inductive approach where we derived each taxonomy scheme via a group read of the corpus. The group read consisted of individually reading mutually exclusive sets of documents, then convening to discuss and determine themes. In developing the impact statement taxonomy, we read a total of 340 statements, which was over 40% of the number of all collected statements. We note that not all of the statements in the group-read stage were in the final because we updated the scraper following the group read. In developing the ethics review taxonomy, we read reviews of 41 of the 51 links for which ethics reviews were made available. We conducted a trial run of coding ethics reviews (corresponding to 20 of the 51 links) and revised the taxonomy to reduce ambiguity in our coding scheme.

Finally, we qualitatively coded impact statements and ethics reviews based on our taxonomy schemes. In particular, we coded 30% randomly-selected impact statements.

3 IMPACT STATEMENT FINDINGS

Below we take stock of discussions of societal impacts in 2021 NeurIPS papers. Compared to analyses of statements from 2020 (e.g., [19]), ours centers more on discussions of negative impacts, which was a directive for NeurIPS authors in 2021. Based on our reading of statements, we find that *Agency* and *Responsibility* are informative attributes by which to study statements. Agency refers to whether authors seem to believe they have control over identification or mitigation of negative societal impacts, and responsibility refers to authors appearing to take ownership over identifying impacts and detailing mitigation strategies. We observe that authors frequently do not express agency and either deny responsibility or assign it to other parties. These observations are based on our read

[5] The 50 papers equated to 51 links because one paper was part of the NeurIPS Consistency Experiment and reviewed twice.

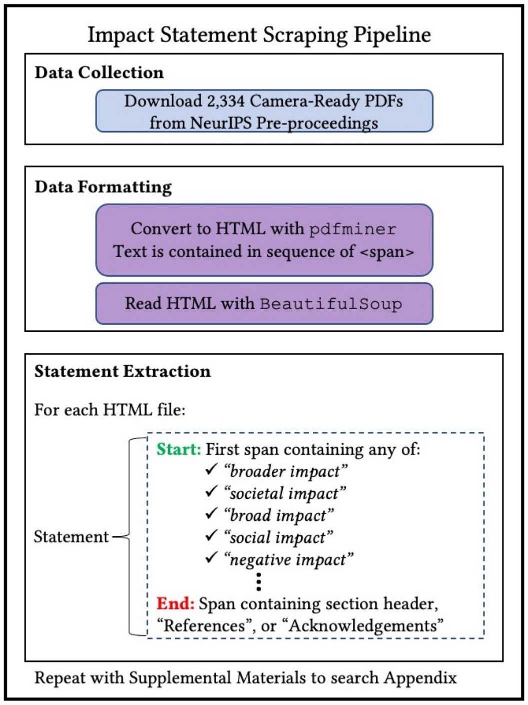
Impact Statement Scraping Pipeline
Data Collection
Download 2,334 Camera-Ready PDFs from NeurIPS Pre-proceedings
Data Formatting
Convert to HTML with pdfminer
Text is contained in sequence of <span>
Read HTML with BeautifulSoup
Statement Extraction
For each HTML file:
Statement
Start: First span containing any of:
✓ "broader impact"
✓ "societal impact"
✓ "broad impact"
✓ "social impact"
✓ "negative impact"
⋮
End: Span containing section header, "References", or "Acknowledgements"
Repeat with Supplemental Materials to search Appendix

Figure 2: Since authors were no longer required to discuss societal impacts in a dedicated "broader impacts" section in 2021, we designed a scraping pipeline that flexibly locates societal-impact discussions, if present. After downloading the main paper and supplemental materials from the NeurIPS Pre-Proceedings, we convert each PDF to HTML. We then define the impact statement as all text between the key phrase-conditional start and end elements. We defined the start and end elements based on our manual examination of papers and validated the entire pipeline by checking for false negatives.

of a random sample of 231 statements, 30% of all the statements we scraped.

3.1 Agency

We find two mechanisms by which authors seem to express a lack of agency in identifying or mitigating negative societal impacts.

Adversarial Users. Authors write about ways their work could be used by "adversarial," "malicious," or "nefarious" actors (14%, $N = 33$). For instance, the bad actor could use generative adversarial networks (GANs) to create deepfakes [43],[‡] conduct image surveillance [39],[‡] or manipulate people [51].[‡] In this way, authors seem to be conveying a perception that certain negative outcomes could be inevitable given an ill-intentioned user.

In discussing adversarial use cases, authors often frame their contributions as double-edged swords. While efficiency and control are often useful features, authors sometimes note that these same capabilities can bolster malicious efforts. For instance, we see an example of this dual impact theme in [63]:[‡]

> Further, verification exposes flaws in neural network models, which on the one hand can help improve their robustness, but on the other hand, can be exploited by an adversary.

Among discussions of malicious use cases, we observe that authors write about common themes of use cases, such as generating deepfakes and fake news and conducting privacy-invasive surveillance. We also notice that the severity of the repeated malicious use cases differ widely. For instance, in addition to the abovementioned use cases, authors also state that their work could aid the development of biochemical weapons [50].[‡] In sum, there is a small list of negative impacts that are repeatedly cited, but these impacts differ widely in severity.

Improper Input or Application. We find that authors write about negative societal impacts resulting from improper input to a model or improper application. Regarding input, authors sometimes write that negative impacts would arise if biased data are provided to the model or if the model builds on top of an existing biased parent model (14%, $N = 23$). For example, instances of both of these identifications of causes appear in [42][‡] and [30]:[‡]

> However, such learning is only as good as the data used for training, and if the data is not unbiased, this could lead to significant issues related to fairness and could also lead to societally undesirable outcomes.

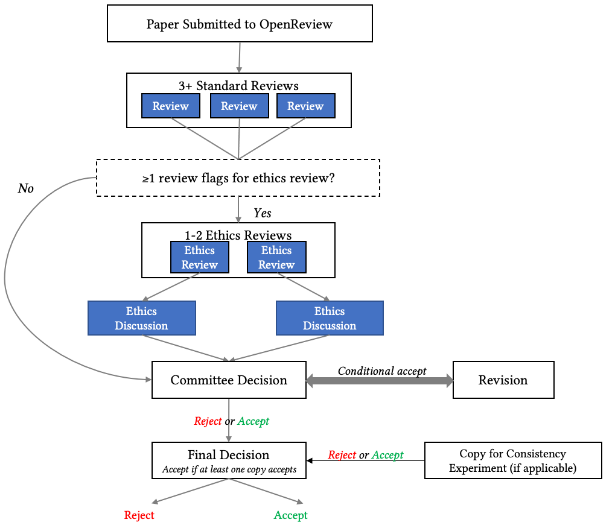

Figure 3: The above figure summarizes the NeurIPS review process to contextualize the role of the ethics reviews. Ethics reviews are conducted if at least one original reviewer flags the paper. For this reason, authors will have already received the scores from their standard reviews by the time they review their ethics reviews. Further, only in cases of conditional acceptance are recommendations from the ethics review verified and accounted for prior to acceptance.

[‡] Reference is part of our NeurIPS 2021 impact statement and ethics review dataset and can be found in the appendix.

> Finally, we bear in mind that—as with any other imitation learning method that aims to match the expert's policy—[Invariant Causal Imitation Learning] can have potential negative societal impacts if the expert's policy is flawed in the first place.

In discussions of deployment, authors frequently define the settings under which their models are meant to perform, and identify potential negative impacts that could arise outside the expected settings; in the NeurIPS ethics guidelines [5] these were termed "application" dependent negative societal impacts. These directives included limitations to model capabilities, warnings for applying the contribution to certain tasks or domains, and discussions of forms of misuse (18%, $N = 41$). For example, [48]‡ states that the proposed image denoising model may not work as expected if a distribution assumption is violated:

> In real environments, the prior knowledge of noise distribution may not be available, and the noise model could not be modeled by exponential family noises.

[33]‡ more explicitly notes that decisions outside the authors' control could lead to negative impacts:

> Unfortunately, like many advances in deep learning for videos, this approach can be utilized for a variety of purposes beyond our control.

By placing the threat of a negative impact in a source outside what is implied to be the core contribution of the paper—often a model, algorithm, or system—and rather in improper use or input, it seems that authors imply a lack of agency in mitigating negative societal consequences. It also seems that authors sometimes appear to struggle with anticipating potential negative societal impacts given the number of contextual factors outside their control.

3.2 Denying Responsibility

Separate from expressing a lack of agency, we observe that authors sometimes also deny the need to take responsibility in the first place. We see this occur both explicitly in the form of minimizing negative societal impacts and implicitly in the form of authors infrequently proposing mitigation strategies.

Minimizing Negative Impact. Prior work analyzing 2020 NeurIPS broader impact statements found that authors often found ways to deflect or downplay the need to discuss negative societal impacts, effectively removing their participation [3]. In 2021's statements we find that authors write that theoretical work does not have societal impact (7.8%, $N = 18$). For example, [65]‡ writes:

> Finally, as a theoretical work, we do not anticipate any potential negative societal impacts of our paper.

In addition, we find that despite the directive to focus on negative impacts, many papers focus, and some exclusively, on positive societal impacts (36%, $N = 84$), thus minimizing emphasis on negative impacts. The vast majority of papers that discuss positive impacts juxtapose positive impacts next to negatives impacts. In a small subset of papers (1%, $N = 2$), authors claim that the positive impacts outweigh the negative ones, such as in the case of [46]:‡

> Extra caution should be taken that these methods are steered away from applications that could be used maliciously, but we believe that, ultimately, these AI advances will do more good than harm.

Lack of Mitigation. Though the NeurIPS ethics guidelines [5] ask authors to discuss mitigation strategies for any identified negative impacts, in practice, we find that statements rarely discuss mitigation. We divide statements based on those that propose a mitigation strategy (11%, $N = 26$) and those that implement one (2.2%, $N = 5$). For papers that discuss mitigation, we observe that mitigation strategies differ along two dimensions: granularity and immediacy. Regarding granularity, some mitigation strategies proposed concrete action items to address negative impacts stemming from the paper's contributions while others suggest more wide-reaching research agendas to tackle negative impacts from the paper along with associated issues. An example of a concrete mitigation approach is the series of recommendations listed in [58]‡ to reduce the risk of a privacy-violating attack:

> Although the risk could be mitigated to some degree with the specific settings (e.g. small gradient due to pre-trained backbone, deeper network, more pooling layer, a mixture of multiple tasks), the privacy problem should not be ignored since we aim to use this method in collaboration with hospitals where the patient privacy is a matter of the highest priority.

In contrast, [52]‡ presents an overarching mitigation directive to avoid all negative social impacts:

> However, there is also a potential risk that the proposed algorithm could be used as a tool to identify minorities and discriminate against them. It should be ensured that the proposed method cannot be used for any purpose that may have negative social impacts.

By immediacy, we refer to the fact that authors differ in the timeline for implementing mitigation strategies. In a small number of cases, authors implement the proposed mitigation strategy (2.2%, $N = 5$). For example, [60]‡ provides an approximation algorithm to combat environmental costs:

> One negative impact of this research is the significant environmental impact associated with training transformers, which are large and compute-expensive models... To mitigate this, we proposed an approximation algorithm with linear complexity that greatly reduces the computational requirements.

On the other hand, other strategies are much more forward-looking without a clear timeline. [66]‡ provides a mitigation strategy for issues stemming from human-AI misalignment that is implemented in future deployments:

> One potential strategy for mitigating these risks is the use of human preference data... Such data could be used to fine-tune and filter trained agents before deployment, encouraging better alignment with human values.

3.3 Assigning Responsibility

We notice that authors sometimes assign responsibility for identifying or mitigating negative societal impacts in their statements.

In particular, we observe three parties to which authors assign responsibility: practitioners, other authors in a particular subfield, and the broader research community.

Practitioners. Practitioners are often referred to as domain experts or researchers with knowledge of the application area. In these cases, authors argue that it is difficult to predict the negative societal impacts without knowledge of the specific domain, so authors leave identification and mitigation to the user. An example of such a delegation to the practitioner occurs in [59]:[‡]

> However, in spite of the potential positive aspects, practitioners need to pay sufficient attention to various perspectives on their problems and our assumptions when trying to apply our proposed method to real-world problems.

When responsibility is passed to the practitioner authors often use words such as "caution" or "care," the task assigned to the practitioner is large and undefined, and little further guidance is provided.

Subfield. A common argument in impact statements is that the paper shares the same negative societal impacts as other papers in a particular subfield or research area. The argument is that the paper is one instance of a larger class so any issues applying to the parent class are inherited in the paper. These claims often carry two implications. First, these claims serve as a defense of the paper, arguing that the paper does not introduce any additional negative societal impacts and only perpetuates existing negative impacts. Second, these claims are also used to more explicitly remove responsibility for mitigating the negative impacts. Authors express difficulty disentangling the issues in their paper from those in the community. The implied result is that the issues are beyond the scope of the paper and cannot be mitigated within the individual paper. An example of assigning responsibility to the subfield occurs in [31]:[‡]

> Clustering methods in general have potential issues with fairness and privacy, which applies also to our work, but our research is not expected to introduce new negative societal impact beyond what is already known.

Future Work. The final form of delegating responsibility we identify is assigning identification and mitigation of negative impacts to future work. An example of pushing mitigation to future work occurs in [32]:[‡]

> Careful consideration of the intended application, and further study of uncertainty quantification in meta-learning approaches will be essential in order to minimize any negative societal consequences of [the paper's contribution] if deployed in real-world applications.

We interpret the use of recommending future work within the impact statements as assigning responsibility to the broader research community.

4 ETHICS REVIEW FINDINGS

We consider ethics reviews holistically; that is, we consider not only the content of ethics reviews, but also the original reviews where papers were flagged for ethics reviews and responses by authors to ethics reviews. We characterize ethics reviews by 1) ethical issues raised in the ethics reviews, 2) recommendations ethics reviewers make to authors in light of the identified ethical issues, and 3) the nature of the interaction between the original reviewers, ethics reviewers, and authors.

4.1 Ethical Issues

Here we focus on ethical issues that are raised or discussed by ethics reviewers. Broadly, we find that these issues vary in terms of the category of ethical issue raised ("Category") and the extent to which the issue is expressed by ethics reviewers to be unique to the paper at hand versus an issue that arises more generally in papers of a specific subfield or type of work ("Scope").

4.1.1 Category. Ethics reviews focus on issues falling into two main categories: policy violations and non-policy issues that tend to require more involved ethical deliberation.

Policy. Policy issues ($21.9\%, N = 21$) are raised when ethics reviewers call for further documentation around IRB protocols or approvals, sometimes referring explicitly to the NeurIPS paper checklist item on IRB documentation [21]. For example, ethics reviewer igBF[6] on [62][‡] notes that the paper requires more details on protocols followed for using animal data in experiments:

> The key issue here is the use of this animal data in the experiments. The paper says, "All data coming from monkeys complied with the approved protocol of local authorities", which one of the reviewers pointed out. I think this statement is not sufficient to justify the overall protocol in this work, and additional detail needs to be provided . . . Lastly, the IRB portion, which I presume the data collection study went through is answered with N/A. Having said all of the above, the paper does mention that protocols were observed, but it just doesn't provide enough information on this front.

Ethics reviewer b7jJ raises similar issues about documentation around protocols in [44],[‡] but regarding human subjects research:

> This work involves human subjects. However, the authors do not provide information on whether the experimentation was reviewed and approved by the relevant oversight board. Neither do they provide information involving the humans in the experiment.

We also find that sometimes concerns about plagiarism are acknowledged in ethics reviews, although are also sometimes distinguished from ethical issues that ethics reviews are meant to address. As ethics reviewer enQy writes in their review of [69],[‡] "the plagiarism concern raised by [an original reviewer] is a potential code of conduct violation separate from the NeurIPS ethical guidelines." Along the same lines, ethics reviewer M5uE in a review for [72][‡] calls "plagiarism concerns . . . outside the scope of ethics review."

Last, we find that ethics reviewers point out lack of impact statements or discussion of negative societal consequences, sometimes citing NeurIPS' guidelines. Ethics reviewer GdrQ in a review for [61][‡] explicitly references the call for papers [20]:

[6] We refer to ethics reviewers by their anonymized IDs on OpenReview.

> This work largely ignored the possible negative societal implications of their work, despite the [call for papers] explicitly asking authors to reflect on these for their work.

Non-Policy Issues. Non-policy issues (69.8%, $N = 67$) discussed by ethics reviewers fall into several themes. We find a theme of ethics reviewers discussing the implications of research in terms of how it may perpetuate societal biases or discrimination (e.g., racial bias and discrimination). For example, ethics reviewer YrQ8 for [64]‡ writes that "a future technology which relied on the advances described in this paper could be used to create an exclusionary online atmosphere or promote racially-biased standards of beauty." In reviewing a paper [53]‡ on voice conversion technology, ethics reviewer tLu4 writes that the technology might "provide tools for white supremacists to racially abuse and harass individuals," and in a review for [54],‡ ethics reviewer vxyR writes that "it is not clear which biases the model perpetuates and amplifies: the datasets are not balanced across demographics (gender, age) and they might have worse results for specific slices of populations."

Ethics reviewers also discuss concerns around the use of ethically dubious datasets. Ethics reviewer RMnL in a review for [55]‡ mentions that the Penn94 dataset [24] "was controversial when produced (raising privacy and other concerns)," and indicates that this should have been mentioned by the paper authors, who ostensibly used the data in their work. Other datasets mentioned due to ethical concerns include the CIFAR-10 benchmark dataset [16] (e.g., discussed by ethics reviewer 6ELm in reviewing [45]‡) and the 80mTiny dataset[7] [28] (e.g., discussed by ethics reviewer 1QET in reviewing [57]‡).

Ethics reviewers also discuss concerns around privacy and surveillance. For instance, ethics reviewer 7ABz writes in a review of [38]‡ that "it would be useful to think more fully about what it means to infer higher order intrinsic characteristics and how such inferences might both leak private (or at least withheld) information." Ethics reviewer hTQw in reviewing [56]‡ also raises privacy concerns, pointing out that "the ability to conduct few shot identification of individuals based on their decisions is likely to be of interest to marketers, law enforcement, and many other entities in ways that might give rise to privacy violations and abuse." We also find that ethics reviewers sometimes discuss concerns around the research's impact on the environment (e.g., ethics reviewer kHQK mentions in their review of [40]‡ that "ethical considerations relevant to this work are the environmental and financial costs associated with developing and deploying large models").

4.1.2 Scope. We find that ethics reviewers express the idea that an ethical issue that applies to the paper being reviewed is part of a larger issue arising in its subfield or field overall (14.6%, $N = 14$). In this way we find that ethics reviewers make nods to the "scope" of the issue. Ethics reviewer YrQ8 writes the following about [64]:‡

> This paper does raise ethical issues in my opinion. As best I can tell, these issues are entirely associated with the general topic of this paper, conditional generative models of high-dimensional images, as opposed to the specific algorithmic advances proposed in the paper. However, by advancing the state of the art in this sub-field, this work inherits the ethical issues of the broader problem.

At times, ethics reviewers note that the paper's ethical issues arise more generally in the topic area to justify why the paper should not be rejected on ethical grounds. For example, ethics reviewer TQwK describes issues of misuse regarding [36]:‡

> The authors acknowledge that [Conditional Generative Adversarial Networks (cGANs)] could be misused with malicious intent so there is some ethical considerations when improving state of the art. This doesn't seem particularly unique to cGANs and in my opinion this type of concern does not preclude acceptance . . .

Ethics reviewer 17Tc invokes a similar argument for why [23]‡ should not be rejected for using potentially ethically questionable datasets (JFT-300M [27] and JFT-3B), writing that they "are reasonably established datasets and the question of their use should not be assigned to a single set of authors." In this way, the ethics reviewer makes an explicit statement about the extent to which authors are responsible for addressing or accounting for certain ethical issues that may be part of a larger pattern.

Sometimes ethics reviewers note that a particular ethical issue may be an instance of a larger problem, and indicate either explicitly or implicitly that the paper authors must still engage with the issue more deeply than they already have in their original paper. Ethics reviewer k7bd takes this approach in their review on [67]‡:

> The authors do acknowledge that their approach is 'data hungry' (as pointed out by one of the reviewers). But their response to this concern is, basically, 'so is every other modeling approach.' This seems like an important first step but does not seem to address the core issues raised above . . .

Similarly, ethics reviewer iUg6 writes about [58]‡ that "[t]he ethics issues that arise are not a result of the instant research, but with systems of federated and split learning themselves," and further recommends that the authors write more specifically about how "criticisms of [federated learning] with regard to privacy" apply to the specific work particularly because of its use of sensitive health data.

4.2 Ethics Reviewer Recommendations

Ethics reviews contain a dedicated section for ethics reviewers to make recommendations for actions paper authors should take to address the ethical issues raised. We consider recommendations that ethics reviewers write in these sections and elsewhere in their reviews. We find that recommendations vary in terms of the extent of actions authors are suggested to take (in terms of identifying an issue or mitigating it) and in terms of the nature of recommended changes (i.e., adding or removing content).

4.2.1 Extent of Recommended Actions. Ethics reviewers often make recommendations around identifying, specifying, or discussing an ethical issue while at times they also make recommendations that are more actively oriented toward mitigating downstream harm. We note that some ethics reviewers make recommendations around both identification and mitigation.

[7] Note on the dataset's retraction: https://groups.csail.mit.edu/vision/TinyImages/

Identification. Ethics reviewers often make recommendations around identification of ethical issues (51%, $N = 49$). Often this means further discussing a potential negative consequence or ethical issues discussed in the review process. For example, ethics reviewer k7bd, when reviewing [67],[‡] notes that the authors should take care to add discussion about certain ethical issues but need not resolve them entirely, in this way drawing a distinction between identification and mitigation:

> It is possible for the researchers to address the [two] ethical concerns raised above. They do not have to 'fix' them but they should be able to speak to how these concerns factor into their sense of the value of their proposed model.

Ethics reviewer VZYf of [61][‡] writes that despite the challenges of considering downstream consequences of "upstream" work, the paper would benefit from more discussion of societal impacts:

> This is a very 'upstream' methodological contribution, rather than more applied future work in which these concerns would arise more intuitively. Nevertheless, the paper would have benefited from some discussion of the possible applications of this methodology, and the societal implications of those applications.

Mitigation. Ethics reviewers sometimes recommend that authors take steps more oriented toward mitigating negative consequences or downstream harms, though recommendations toward mitigation (27.1%, $N = 26$) are rarer than recommendations around identification. At times this means including a more in-depth discussion of methods of mitigation. In a review of [47],[‡] ethics reviewer 2Zj9 points out that actually implementing a mitigation strategy is ideal, but that instead, discussion of these strategies can strengthen the paper:

> Ideally, the research would have a defense to this privacy attack available at the same time that the issue is demonstrated. Short of that, a more robust discussion of potential protection mechanisms, eg whether current methods such as [differential privacy] might help, would be useful.

Similarly, ethics reviewer pVfn in their review of [49][‡] notes that the authors should add specific "recommendations for how researchers in this space can responsibly conduct their work (choice of datasets, protocol for sharing models, etc.)" in light of surveillance concerns. In this way, the recommendation goes beyond identifying issues and urges authors to make steps toward mitigating these issues, in this case by outlining how other researchers may responsibly conduct future work.

Other times, recommendations oriented toward mitigation extend beyond adding more discussion. For example, in a review for [55],[‡] a paper which introduces new datasets, ethics reviewer RMnL asks the authors to add more documentation around the new datasets:

> Emerging best practices require data sets to be accompanied by documentation that supports sound and ethical use.The addendum provides very limited information about the existing datasets and makes no effort to assess whether they are appropriate to use in benchmarking or other research.

In a review of [37],[‡] ethics reviewer pEAe also recommends more documentation, specifically around "the performance of the method on different subjects in the dataset of human faces, instead of only on the entire dataset or the chosen ones that are not representative for all kinds of subjects."

4.2.2 *Nature of Changes.*

Ethics reviewers tend to recommend adding content as opposed to removing content as part of recommended changes.

Addition. Ethics reviewers often recommend adding material to papers (66.7%, $N = 64$). Sometimes this takes the form of asking authors to provide additional documentation (e.g., ethics reviewer b7jJ's recommendation to add more documentation on protocols around human subjects research as seen in [44][‡]). Other times, recommended additions focus more on adding further discussion around ethical issues or other details. Ethics reviewer Ws2D, for example, suggests that the authors of [54][‡] "expand the limitations discussions as well as add more discussion of potential misuses" and "add more details about the data sources" while ethics reviewer Ag79 writes in a review of [29][‡] that "[t]he authors should acknowledge that being able to recover membership of an individual in a disadvantaged group may itself be highly problematic." As an example of an instance where the recommended addition is more general, ethics reviewer 1CRT in a review of [41][‡] writes that in light of a missing impact statement, the authors "can add a discussion of broader impacts to the Conclusion section without substantive changes to the rest of the paper."

Removal. Although less frequent than additions, ethics reviewers sometimes suggest that authors remove content from their papers (6.3%, $N = 6$). For instance, ethics reviewer 7ABz implies that the authors of [38][‡] can remove genders in their dating scenario, writing that "there is no need to specify the genders of the potential daters." In a review of [68],[‡] ethics reviewer mpwL recommends that the authors remove an example attack from the paper, while ethics reviewer 1QET recommends that the authors of [57][‡] remove the use of an ethically questionable dataset (80mTiny [28]). Ethics reviewer mu5w recommends that the authors of [70][‡] remove experiments that raise ethical questions:

> Given that the question of whether someone is attractive or not is not a question devoid of politics, I would recommend that the authors simply remove the experiments predicting attractiveness or replace them with something that is less overtly fraught with issues.

4.3 Interaction Between Authors, Original Reviewers, and Ethics Reviewers

We characterize interactions between parties involved in the ethics review process (authors, original reviewers, ethics reviewers) in terms of the extent to which authors appear to accept suggestions by ethics reviewers, modes of justifying the work used by authors when responding to ethics reviewers, and level of consistency between ethical issues flagged in original reviews and discussions in ethics reviews.

4.3.1 Acceptance of Suggestions. We find that when applicable, there is always an openness on the part of authors to adopting some or all suggestions from ethics reviewers (58.3%, $N = 56$) when authors respond to ethics reviews (which is not always the case). We often find that authors accept suggestions, at least in some capacity, offered by ethics reviewers and respond by stating that they will attempt to make the suggested changes or describing the changes they plan to make in response to ethics reviews.

Sometimes responses contain a mix of refuting claims by ethics reviewers and accepting suggestions. For example, ethics reviewer GdrQ in their review of [61]‡ recommends that the authors include a discussion of negative societal impacts. In their response, the authors appear to refute the ethics reviewer's claim that the paper does not acknowledge ethical issues while also committing to making certain paper changes (e.g., "[w]e will expand this discussion to include specific applications in social robotics . . . and surveillance, to broaden the scope to societal implications"). Sometimes authors are more in agreement with ethics reviewers. As the authors of [35]‡ write in response to their ethics reviewers, "we are in complete agreement that the dangers of designing algorithms for model extraction are very real" and go on to note that they "will make sure in the final version of the draft to incorporate elements of this discussion so that this point becomes clear and also formulate new directions towards mitigating such attacks."

4.3.2 Justification Mechanisms. We find that at times authors appear to justify their work in different ways in their responses to ethics reviews. For example, we find that sometimes authors employ a type of ***Citing*** mechanism (10.7%, $N = 6$),[8] where they cite other work to help explain certain choices in their research. In agreeing to make a change in the dataset used in the paper, the authors of [57]‡ also provide an explanation for their original dataset choice:

> We will make sure to remove the 80mTiny dataset and substitute it with another dataset . . . We used the 80mTiny dataset simply because OAT (published in ICLR 2021) had used it.

Authors also sometimes appear to use a ***Limiting*** mechanism (21.4%, $N = 12$) in responses, essentially limiting the scope of scenarios where their work might apply. For example, the authors of [71]‡ write that "while [they] hope [their] model can be used by the community for research, [they] actively do not want the model to be used for production purposes." Authors also use a ***Correcting*** mechanism (19.6%, $N = 11$), where they appear to correct a statement or idea expressed by ethics reviewers. Last, we note that authors at times mention or describe ***Positive Consequences*** (26.8%, $N = 15$) in their responses, which sometimes function as a way of justifying the work at hand. For example, in response to ethics reviewer 8nNQ, the authors of [49]‡ write the following:

> Besides the possible negative effects, we highlight that there are also many positive societal effects. The technology can be applied in 1) video conferencing in a silent or crowded environment, 2) audio enhancement using visual information, 3) conversation in a long-distance, and 4) conversation with people who cannot make a voice.

4.3.3 Consistency (between Original Reviews and Ethics Reviews). We find that there is often at least some consistency between ethical issues flagged by original reviewers and ethical issues addressed or discussed by ethics reviewers (89.6%, $N = 86$). (Note that if the original reviewers flag papers for ethics reviews but do not specify why, we automatically consider ethics reviews to be consistent with the original reviews, and as long as ethics reviews mention at least one of the flagged issues, we also consider the ethics review to be consistent with the original reviews.) As such, ethical issues pointed out by original reviewers are often considered by ethics reviewers; that is, there seems to usually be effective communication between original and ethical reviewers. However, there are times where there appears to be inconsistency between why a paper was flagged for ethics review by original reviewers and what ends up being discussed or considered by ethics reviewers (10.4%, $N = 10$). For example, in a review for [34],‡ original reviewer zxcr notes that there may be issues with lack of anonymity of a funding source, but ethics reviewer D5WN does not address this issue (or lack thereof) in their review.

5 RECOMMENDATIONS

We reflect on our analysis to make recommendations for future iterations of impact statements in the artificial intelligence research community. The NeurIPS ethics guidelines [5] set forth several recommendations for what to consider in the ethics reviews and impact statements, such as studying disparate impacts on marginalized communities; however, given the newness of the statement, we posit that further iterations on recommendations for authors in writing impact statements could help deepen ethical deliberations. Hence, below we suggest several potential recommendations for authors writing impact statements.

5.1 Transparency

Further incorporate the impact statement into the research process. We echo previous recommendations [14, 19] to integrate the impact statement into the research process by considering societal impact at the beginning of projects and also encourage authors to draw on lessons from past impact statements. Authors who consider the impact statement earlier on in the research process may, for example, be able to make more significant changes to the research questions than authors who write the statement once the research has already been completed. Similarly, if NeurIPS continues to encourage impact statements, it may be that statements authors write in previous years help inform aspects of future submissions. For example, as past impact statements frequently propose future work or ethical challenges facing subfields, authors can synthesize past statements to motivate research directly addressing these issues.

Identify a wider range of possible negative impacts by considering harms resulting from the actions of well-intentioned parties. We find that impact statements tend to focus on a narrow set of impacts, such as the proliferation of misinformation and surveillance that infringes upon civil rights. Authors may want to consider expanding discussions of potential uses and applications of and for their models to capture a wider range of potential negative consequences. For

[8] Note that if authors provide a common response to multiple ethics reviews, we consider the response as if it were individually responding to each review in obtaining counts. For percentages in the Justification Mechanisms section, the denominator is the number of ethics reviews with author responses.

example, we find that authors sometimes describe harms arising from the actions of bad actors. Focusing on these harms could limit the discussion of broader impacts to harms resulting from misuse by parties who are clearly ill-intentioned. However, a wider range of consequences might be identified if authors consider the negative consequences that could result from a technology's use by well-intentioned parties. Identification of these consequences could also aid in developing mitigation strategies that align with uses both ill- and well-intentioned. For instance, while the mitigation strategies for a bad actor might be to restrict access to the code or data, a strategy for mitigating negative impacts arising from a well-intentioned user's actions might call for increased transparency (e.g., of a model) or a better understanding of its limitations.

Be explicit about the scope of ethical issues. We find that at times, ethics reviewers point out the ways in which an ethical issue raised for a given paper is part of a larger issue that applies more broadly to a type of technology or subfield. Thus, there may be certain classes of technologies with recurring ethical questions or documented examples of harms resulting from their use. We encourage authors to state whether an identified negative impact is specific to an individual paper or applicable to a broader subfield, beginning with issues that are specific to the paper. One inhibition to discussing issues stemming from the subfield may be fear of repetition or being the target of undue blame. We argue that discussing recurring issues will be a move toward the community mitigating negative impacts.

5.2 Accountability

Reflect further on ways in which researcher decisions or actions can help mitigate negative consequences. We find that authors sometimes focus less on their own responsibility around the consequences of their work, and rather ascribe responsibility to other external actors/users, sometimes though not always defined. In this way, authors at times appear to frame their contributions as neutral, for example noting their inability to control the ways in which external actors may deploy their models. However, describing the work as neutral creates a false dichotomy between design and deployment, which could lead to less accountability on the part of researchers. It could also discourage authors from deeper reflection around the ways that implications of their contributions *are* under their control. Thus, the process of writing impact statements where authors describe the ways in which the impacts of their work are primarily outside their control may paradoxically strengthen beliefs among researchers that they are not responsible for, or do not have agency around, mitigating the negative implications of their contributions. While there are almost certainly instances in which authors are unable to foresee or mitigate some downstream consequences, we encourage authors to think defensively about potential negative impacts and corresponding mitigation strategies.

Explicitly weigh the positive impacts of their work with potential harms and qualify their rationale. Several authors claim that the benefits of their work will or are likely to outweigh potential harms, but without conducting a thorough, less subjective cost calculus or weighing of impacts. In other words, if authors argue that positive impacts are likely to outweigh negative impacts, it could be useful to encourage more rationale around this reasoning, including evidence that may bolster such claims. In some cases, authors enumerate potential harms but do not weigh the benefits against the harms, leaving that judgment to the reader and potentially encouraging less responsible, less critical use of the work. Further, if authors point to a substantive potential harm, e.g., development of chemical weapons, but do not clearly lay out why the benefits outweigh that potential harm, it sheds reasonable doubt around the value of the work for society.

Discuss mitigation strategies with more granularity. While the current ethics guidelines [5] include an expectation that authors include a discussion about methods to mitigate enumerated risks, not all authors outline the steps they have taken or will take to mitigate the negative impacts of their work. Further, authors who discuss mitigation tend to delegate it to later work or propose wide-reaching strategies that are difficult to operationalize. In the ethics review process, ethics reviewers also tend to recommend changes focused on identifying ethical issues moreso than changes geared toward mitigation. Thus, further encouraging authors to discuss or implement mitigation strategies could be useful in engaging the artificial intelligence research community in more actively taking part in reducing downstream negative impacts.

5.3 Accessibility

Reduce jargon in impact statements. Impact statements can play a role in cultivating a wider understanding, for instance across subfields in computer science research, around the potential benefits and harms of cutting-edge research. However, we suggest that the extent to which these statements can reach audiences outside specific subfields is limited if statements are filled with overly technical language. Further analysis into the language used in impact statements, e.g., if they tend to mention concepts or use phrases that may not be widely understood outside certain areas of computer science research, could help pinpoint ways in which statements can be written in ways that invite a wider audience, beginning with computer science researchers in other subfields. Making the statements more accessible can also help ethics reviewers, who may be domain experts without technical specialty in the particular subfield, to more comprehensively evaluate papers.

6 CONCLUSION

In this work we qualitatively analyze impact statements and ethics reviews from NeurIPS 2021. We characterize impact statements in terms of how authors assign responsibility and express agency around mitigating negative consequences. In addition, we describe themes in the ethics reviews with regards to issues raised by ethics reviewers, changes they recommend to papers, and interactions among authors, ethics reviewers, and original reviewers in general. Based on our findings, we make recommendations for future impact statements institutionalized by the artificial intelligence research community that could help support researchers in engaging in ethical deliberations regarding their work.

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.*

[2] Carolyn Ashurst, Rosie Campbell, Deborah Raji, Solon Barocas, and Stuart Russell. 2020. Workshop on Navigating the Broader Impacts of AI Research. Neural Information Processing Systems Conference. https://ai-broader-impacts-workshop.github.io.

[3] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements – Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA.

[4] Chloé Bakalar, Sarah Bird, Tiberio Caetano, Edward Felten, Dario Garcia-Garcia, Isabel Kloumann, Finn Lattimore, Sendhil Mullainathan, and D. Sculley. 2018. Workshop on Ethical, Social and Governance Issues in AI. Neural Information Processing Systems Conference. https://sites.google.com/view/aiethicsworkshop/the-workshop?authuser=0.

[5] Samy Bengio, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, Inioluwa Deborah Raji, and Marc'Aurelio Ranzato, Ranzato. 2021. Ethics Guidelines. https://nips.cc/public/EthicsGuidelines

[6] Samy Bengio, Inioluwa Deborah Raji, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 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/

[7] Samy Bengio, Inioluwa Deborah Raji, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. A Retrospective on the NeurIPS 2021 Ethics Review Process. *Medium. https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/* (2021).

[8] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. Introducing the NeurIPS 2021 Paper Checklist. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b.

[9] Lutz Bornmann. 2013. What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 217–233.

[10] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. In *Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems*.

[11] A. Feder Cooper, Benjamin Laufer, Emanuel Moss, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA.

[12] Fei Fang, Joseph Aylett-Bullock, Marc-Antoine Dilhac, Brian Green, natalie saltiel, Dhaval Adjodah, Jack Clark, Sean McGregor, Margaux Luck, Jonnie Penn, Tristan Sylvain, Geneviéve Boucher, Sydney Swaine-Simon, Girmaw Abebe Tadesse, Myriam Côté, Anna Bethke, and Yoshua Bengio. 2019. Joint Workshop on AI For Social Good. Neural Information Processing Systems Conference. https://aiforsocialgood.github.io/neurips2019/.

[13] Thirty fourth Conference on Neural Information Processing Systems. 2020. NeurIPS 2020 Subject Areas. https://nips.cc/Conferences/2020/PaperInformation/SubjectAreas.

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

[15] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, et al. 2021. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *arXiv preprint arXiv:2112.09544* (2021).

[16] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images.* Technical Report. University of Toronto.

[17] HT Lin, MF Balcan, R Hadsell, and MA Razato. 2020. Reviewing is Underway! *Medium. https://medium.com/@NeurIPSConf/reviewing-is-underway-a5532d4615ec* (2020).

[18] Hsuan-Tien Lin, Maria-Florina Balcan, Raia Hadsell, and Marc'Aurelio Ranzato. 2020. Getting Started with NeurIPS 2020. *Medium. https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28* (2020).

[19] 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*. ACM, New York, NY, USA, 795–806.

[20] Neural Information Processing Systems Conference (NeurIPS). 2021. NeurIPS 2021 Call for Papers. https://neurips.cc/Conferences/2021/CallForPapers.

[21] Neural Information Processing Systems Conference (NeurIPS). 2021. NeurIPS 2021 Paper Checklist Guidelines. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist.

[22] Partnership on AI. 2020. Publication Norms for Responsible AI. https://www.partnershiponai.org/case-study/publication-norms/.

[23] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems* 34 (2021), 8583–8595.

[24] Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, 4292–4293. http://networkrepository.com

[25] Igor Rubinov, Risi Kondor, Jack Poulson, Manfred K. Warmuth, Emanuel Moss, and Alexa Hagerty. 2019. Workshop on Minding the Gap: Between Fairness and Ethics. Neural Information Processing Systems Conference. https://mindingthegap.github.io/.

[26] Megan M Skrip. 2015. Crafting and evaluating Broader Impact activities: a theory-based guide for scientists. *Frontiers in Ecology and the Environment* 13, 5 (2015), 273–279.

[27] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*. 843–852.

[28] Antonio Torralba, Rob Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11 (2008), 1958–1970.

APPENDIX: REFERENCES FROM OUR NEURIPS 2021 IMPACT STATEMENT AND ETHICS REVIEW DATASET

[29] Adarsh Barik and Jean Honorio. 2021. Fair Sparse Regression with Clustering: An Invex Relaxation for a Combinatorial Problem. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=l-0rLXvctI

[30] Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. 2021. Invariant Causal Imitation Learning for Generalizable Policies. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=715E7e6j4gU

[31] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. 2021. Coresets for Clustering with Missing Values. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=1H6zA8wIhKk

[32] John F Bronskill, Daniela Massiceti, Massimiliano Patacchiola, Katja Hofmann, Sebastian Nowozin, and Richard E Turner. 2021. Memory Efficient Meta-Learning with Large Images. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=x2pF7Tt_S5u

[33] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. 2021. NeRV: Neural Representations for Videos. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=BbikqBWZTGB

[34] Haibo Chen, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Artistic Style Transfer with Internal-external Learning and Contrastive Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=hm0i-cunzGW

[35] Sitan Chen, Adam Klivans, and Raghu Meka. 2021. Efficiently Learning One Hidden Layer ReLU Networks From Queries. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=oZg-aOyHL-h

[36] Si-An Chen, Chun-Liang Li, and Hsuan-Tien Lin. 2021. A Unified View of cGANs with and without Classifiers. *Advances in Neural Information Processing Systems* 34 (2021).

[37] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. 2021. DIB-R++: Learning to Predict Lighting and Material with a Hybrid Differentiable Renderer. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=aLMEzZnAoPo

[38] Yuzhou Chen, Baris Coskunuzer, and Yulia Gel. 2021. Topological Relational Learning on Graphs. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=YOc9i6-NrQk

[39] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. 2021. CATs: Cost Aggregation Transformers for Visual Correspondence. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=eVuMspr9cu5

[40] Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems* 34 (2021).

[41] Kyra Gan, Su Jia, and Andrew A Li. 2021. Greedy Approximation Algorithms for Active Sequential Hypothesis Testing. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=XOSrNXGp_qJ

[42] Antonious M. Girgis, Deepesh Data, and Suhas Diggavi. 2021. Renyi Differential Privacy of The Subsampled Shuffle Model In Distributed Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=SPrVNsXnGd

[43] Denizalp Goktas and Amy Greenwald. 2021. Convex-Concave Min-Max Stackelberg Games. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=gaftyBQ4Lu

[44] Abhinav Gupta, Marc Lanctot, and Angeliki Lazaridou. 2021. Dynamic population-based meta-learning for multi-agent communication with natural language. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=NFurmj-rIWe

[45] Paul Haider, Benjamin Ellenberger, Laura Kriener, Jakob Jordan, Walter Senn, and Mihai Petrovici. 2021. Latent Equilibrium: Arbitrarily fast computation with arbitrarily slow neurons. *Advances in Neural Information Processing Systems* 34 (2021).

[46] Brandon G Jacques, Zoran Tiganj, Marc Howard, and Per B Sederberg. 2021. DeepSITH: Efficient Learning via Decomposition of What and When Across Time Scales. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=tn6vqNUJaEW

[47] Jinwoo Jeon, Jaechang Kim, Kangwook Lee, Sewoong Oh, and Jungseul Ok. 2021. Gradient Inversion with Generative Image Prior. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=x9jS8pX3dkx

[48] Kwanyoung Kim and Jong Chul Ye. 2021. Noise2Score: Tweedie's Approach to Self-Supervised Image Denoising without Clean Images. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=ZqEUs3sTRU0

[49] Minsu Kim, Joanna Hong, and Yong Man Ro. 2021. Lip to Speech Synthesis with Visual Context Attentional GAN. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=x6z8J_17LP3

[50] Johannes Klicpera, Chandan Yeshwanth, and Stephan Günnemann. 2021. Directional Message Passing on Molecular Graphs via Synthetic Coordinates. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=ZRu0_3azrCd

[51] Cassidy Laidlaw and Stuart Russell. 2021. Uncertain Decisions Facilitate Better Preference Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=sNKpWhzEDWS

[52] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. 2021. ABC: Auxiliary Balanced Classifier for Class-imbalanced Semi-supervised Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=1G6jPa9SKYG

[53] Sang-Hoon Lee, Ji-Hoon Kim, Hyunseung Chung, and Seong-Whan Lee. 2021. VoiceMixer: Adversarial Voice Style Mixup. *Advances in Neural Information Processing Systems* 34 (2021).

[54] Yuan Liang, Weikun Han, Liang Qiu, Chen Wu, Yiting Shao, Kun Wang, and Lei He. 2021. Exploring Forensic Dental Identification with Deep Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=YN4TMf3sv52

[55] Derek Lim, Felix Matthew Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Prasad Bhalerao, and Ser-Nam Lim. 2021. Large Scale Learning on Non-Homophilous Graphs: New Benchmarks and Strong Simple Methods. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=DfGu8WwT0d

[56] Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2021. Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess. *Advances in Neural Information Processing Systems* 34 (2021).

[57] Dongmin Park, Hwanjun Song, MinSeok Kim, and Jae-Gil Lee. 2021. Task-Agnostic Undesirable Feature Deactivation Using Out-of-Distribution Data. *Advances in Neural Information Processing Systems* 34 (2021).

[58] Sangjoon Park, Gwanghyun Kim, Jeongsol Kim, Boah Kim, and Jong Chul Ye. 2021. Federated Split Task-Agnostic Vision Transformer for COVID-19 CXR Diagnosis. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=Ggikq6Tdxch

[59] Seohong Park, Jaekyeom Kim, and Gunhee Kim. 2021. Time Discretization-Invariant Safe Action Repetition for Policy Gradient Methods. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=xNmhYNQruJX

[60] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. 2021. Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=mfQxdSMWOF

[61] Chirag Raman, Hayley Hung, and Marco Loog. 2021. Social Processes: Self-Supervised Forecasting of Nonverbal Cues in Social Conversations. *Advances in Neural Information Processing Systems* 34 (2021).

[62] Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago Cadena, Kelli Restivo, George Denfield, Andreas Tolias, and Fabian Sinz. 2021. Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems* 34 (2021).

[63] Harkirat Singh, M. Pawan Kumar, Philip Torr, and Krishnamurthy Dj Dvijotham. 2021. Overcoming the Convex Barrier for Simplex Inputs. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=JXREUkyHi7u

[64] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. D2C: Diffusion-Decoding Models for Few-Shot Conditional Generation. *Advances in Neural Information Processing Systems* 34 (2021).

[65] Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. 2021. Subquadratic Overparameterization for Shallow Neural Networks. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=NhbFhfM960

[66] DJ Strouse, Kevin R. McKee, Matthew Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with Humans without Human Data. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=79zWncwO2p

[67] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=EI2KOXKdnP

[68] Yixu Wang, Jie Li, Hong Liu, Yan Wang, YONGJIAN WU, Feiyue Huang, and Rongrong Ji. 2021. Black-Box Dissector: Towards Erasing-based Hard-Label Model Stealing Attack. https://openreview.net/forum?id=5jaqt-Hsqir

[69] Zhuo Wang, Wei Zhang, Ning Liu, and Jianyong Wang. 2021. Scalable Rule-Based Representation Learning for Interpretable Classification. *Advances in Neural Information Processing Systems* 34 (2021).

[70] Mike Wu, Noah Goodman, and Stefano Ermon. 2021. Improving Compositionality of Neural Networks by Decoding Representations to Inputs. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=jfd_GB546GJ

[71] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=CRFSrgYtV7m

[72] Yinglun Zhu, Dongruo Zhou, Ruoxi Jiang, Quanquan Gu, Rebecca Willett, and Robert Nowak. 2021. Pure Exploration in Kernel and Neural Bandits. *Advances in Neural Information Processing Systems* 34 (2021).