AI Ethics Statements

Analysis and Lessons Learnt from NeurIPS Broader Impact Statements

Carolyn Ashurst*
cashurst@turing.ac.uk
Alan Turing Institute
UK

Emmie Hine*
emmie.e.hine@gmail.com
Oxford Internet Institute
UK

Paul Sedille*
paul.sedille@gmail.com
Harvard Kennedy School
USA

Alexis Carlier
alexis.carlier@governance.ai
Centre for the Governance of AI
UK

ABSTRACT

Ethics statements have been proposed as a mechanism to increase transparency and promote reflection on the societal impacts of published research. In 2020, the machine learning (ML) conference NeurIPS broke new ground by requiring that all papers include a broader impact statement. This requirement was removed in 2021, in favour of a checklist approach. The 2020 statements therefore provide a unique opportunity to learn from the broader impact experiment: to investigate the benefits and challenges of this and similar governance mechanisms, as well as providing an insight into how ML researchers think about the societal impacts of their own work. Such learning is needed as NeurIPS and other venues continue to question and adapt their policies. To enable this, we have created a dataset containing the impact statements from all NeurIPS 2020 papers, along with additional information such as affiliation type, location and subject area, and a simple visualisation tool for exploration. We also provide an initial quantitative analysis of the dataset, covering representation, engagement, common themes, and willingness to discuss potential harms alongside benefits. We investigate how these vary by geography, affiliation type and subject area. Drawing on these findings, we discuss the potential benefits and negative outcomes of ethics statement requirements, and their possible causes and associated challenges. These lead us to several lessons to be learnt from the 2020 requirement: (i) the importance of creating the right incentives, (ii) the need for clear expectations and guidance, and (iii) the importance of transparency and constructive deliberation. We encourage other researchers to use our dataset to provide additional analysis, to further our understanding of how researchers responded to this requirement, and to investigate the benefits and challenges of this and related mechanisms.

*These authors contributed equally to this research.

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

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533780

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

ethics statements,broader impacts,research ethics,conference policies,AI governance,NeurIPS policies

ACM Reference Format:
Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533780

1 INTRODUCTION

In response to the increasing recognition of harms resulting both from deployed systems [5, 7, 18, 43, 45, 46, 57] and research outputs [6, 13, 15, 26], some have called on the machine learning (ML) research community, including ML conferences and journals, to do more to promote ethical research [25]. This has led to a range of initiatives such as codes of ethics [22], ethics committees [16] and ethics review boards [8].

Perhaps the most significant change relating to the publication of papers was introduced by the Neural Information Processing Systems (NeurIPS) conference. The program chairs of NeurIPS, one of the most prominent ML-related conferences, announced that all authors submitting to the 2020 conference must include a broader impact section, in which authors should discuss "the broader impact of their work, including possible societal consequences – both positive and negative" [36, 38]. In addition, a new ethics review process was incorporated into the peer review process. Technical reviewers could flag papers for potential ethical concerns, to be reviewed by a team of ethics experts [30]. These initiatives provoked both praise and criticism, showing a lack of consensus on whether these mechanisms should be adopted [25, 29, 47], and how they should be operationalised [50].

A year later, in their first blog post for the 2021 conference, the 2021 chairs described how they had reviewed the broader impact requirement, taking into account an author survey, views expressed at the Broader Impacts Workshop [4], similar efforts in other communities, and discussions with the ML community and beyond [9]. The chairs stated that "authors want both more guidance around

how to perform machine learning research responsibly and more flexibility in how they discuss this in their papers". In light of this, the chairs removed the requirement to include a separate broader impact section, and introduced a new checklist for authors to include in their submitted paper [39]. The checklist questions relate to a range of responsible research practices, including reproducibility and scientific best practice. Regarding societal impacts, the authors are asked "Did you discuss any potential negative societal impacts of your work?", and are offered some high level guidance on what to consider. While it is no longer mandatory to include a separate impact statement, the guidance suggests that a discussion of negative societal impacts is expected of researchers, though how to incorporate this into their paper is left to the authors [40].

The 2020 broader impact statements thus represent a unique opportunity to investigate the benefits and challenges associated with ethics statement requirements. Such work can help inform decisions around ethics statements in other contexts including funding applications, organisational approval, or for other publication venues considering introducing such requirements, or who have already done so (such as EMNLP [17]). Lessons may also be applied to the NeurIPS checklist approach (since a discussion of negative impacts remains a component), and future NeurIPS policy.

1.1 Summary of findings

Our ultimate goal is to better understand the implications of impact statement requirements and related mechanisms, including their benefits, risks and challenges. As a step towards this goal, our main contributions are: (i) an open source dataset of impact statements from all NeurIPS 2020 papers, along with additional information such as affiliation type, location and subject area [21], (ii) a simple visualisation tool for exploration of the dataset [53], (iii) an initial quantitative analysis of the dataset, in which we explore representation, engagement, themes and valence (§ 3) and (iv) a discussion of benefits, risks and challenges evidenced by these findings, from which we draw several lessons learnt (§ 4).

Based on our analysis, our main findings are as follows.

Voices represented. There is concentration of authors of papers, and therefore impact statements, along both geographic (§ 3.1.3) and institutional (§ 3.1.2) lines. Authors are concentrated in North America (67% of papers have at least one North American affiliation), followed by Europe (29%) and Asia (27%), and a significant number of affiliations are concentrated in a handful of institutions, such as Google (13% have at least one Google affiliation), Stanford (7%) and Microsoft (5%). We also note the large overlap between industry and academic affiliations, with around a third of papers having affiliations from both industry and academia (§ 3.1.1).

Engagement with broader impacts. We find high variation in engagement as measured by statement length (§ 3.2.2) and opt-out rates (§ 3.2.1). While the average statement length is only 169 words (around 7 sentences), the distribution of lengths has a long tail – the longest statement containing over 4000 words. Around 10% of papers choose to effectively opt out of writing a statement, for example by stating that it is "not applicable". These vary greatly by subject area; *Theory* and *Optimization* had the highest opt-out rates (25% and 24%), and *Applications* and *Social Aspects of ML* had the lowest (2% and 1%).

Themes. We find evidence that certain well established topics were common to many impact statements, including privacy, fairness, robustness and safety (§ 3.3). The most frequent words associated with application settings were *medical*, *robots* and *science*.

Valence. We find evidence that authors tend to discuss more positive aspects compared to negative aspects in their statements (§ 3.4). On average, statements included synonyms of *positive* and *strength* 4.6 and 1.3 times respectively, with synonyms of *negative* and *limitation* occurring 3.6 and 0.6 times.

These findings highlight some of the challenges and issues associated with broader impact requirements, such as incentives to downplay negative impacts, lack of incentives to engage deeply with the task, and a concentration of perspectives along geographic and institutional lines. Despite these challenges, our findings also highlight several benefits of such requirements. Some authors did engage with the task thoughtfully, with some taking the opportunity to thoroughly investigate potential ethical issues, indicating that the requirement can promote reflection and awareness raising. The statements also give us a sense of which issues are widely recognised by authors (such as privacy and fairness), which can help us understand which issues are comparatively neglected. Given this was the first year of the requirement, with lightweight guidance and few examples for authors to refer to, we cannot judge its full potential. Even so, there are several lessons we can draw from this initial attempt. These include (i) the importance of creating the right incentives, (ii) the need for clear expectations and guidance, and (iii) the importance of transparency and constructive deliberation. As the new checklist requirement for NeurIPS 2021 papers asks authors whether they have considered potential negative societal impacts, the community should continue to monitor how researchers respond, and to reflect on the utility of such requirements.

1.2 Related work

This ethical review builds on the spirit of other ethics document analyses, including an analysis of NSF data management plans [10] and a review of literature on the ethics of AI in healthcare [32]. In addition, several other efforts have engaged with NeurIPS broader impact statements. Before the final statements were available, work to understand how researchers had responded to the NeurIPS impact statement requirement included an analysis of preprints available before the conference [11], a survey of researcher attitudes towards the requirement [1], and reflections on potential benefits and challenges based on related mechanisms [50]. To date the most significant investigation into the final statements is provided by Nanayakkara et al. [34], who provide a qualitative thematic analysis of a sample of 300 statements. They identify several themes related to how consequences are expressed (such as valence, specificity and uncertainty), the areas of impacts expressed (such as privacy, labor, the environment, efficiency and robustness), and researcher's recommendations for mitigations. Our work complements theirs by providing a dataset and analysis of all 1898 statements with

additional information (such as affiliation location and type), code and visualisation tools for further analysis, and an initial analysis that both builds on some of their identified themes (such as valence and areas of impacts), and asks complementary questions (such as those around engagement, and investigating how engagement, themes and valence differ by subject area, geography and affiliation type). We also discuss evidence of benefits, negative outcomes, causes, and challenges from these findings, in order to summarise recommendations for future self-governance mechanisms.

2 DATA

We obtained the manuscript pdfs for accepted papers from the NeurIPS 2020 proceedings website [35], which contains 1898 papers.[1] We converted the pdfs to XML and extracted the title and impact statement section [42]. We appended this dataset with information about paper subject area, author names, affiliations, affiliation type and affiliation institution locations, as follows. Primary and secondary subject area, as selected by authors on submission, were supplied to us by the NeurIPS programme chairs [48]. Author names [35] and affiliations [28] were obtained from separate scrapes of the NeurIPS papers. Each affiliation was tagged with a location and type (industry or academia) based on [27] and [2] respectively. Further details on dataset generation, and the assumptions and limitations of our dataset can be found in our code documentation [44]. The resulting dataset and our code can be found in the github repository [21]. We also created a simple visualisation tool, for convenient exploration of the dataset, which can be found on Flourish [53].

While our dataset identifies individual NeurIPS authors and their institutions, we do not expect significant ramifications for them from either our dataset or analysis. Paper-related information in the dataset was previously disclosed by the authors or conference, while calculated statement lengths do not reveal anything about content or quality. Our analysis is primarily of larger trends and should not be taken as a comment on individual researchers.

3 ANALYSIS

We divide our analysis into four categories: representation, engagement, themes and valence.

3.1 Representation

In this section we explore whose voices are represented in the 2020 broader impact statements, by investigating the affiliation types, organisations, geography and subject area of the paper authors.

3.1.1 Affiliation type. We classified each affiliation listed as academia or industry, based on [2]. Papers were then categorised as *academia* (if all affiliations were from academia), *industry* (if all affiliations were from industry), or *mixed* (if the paper had both academic and industry affiliations listed). According to this categorisation, 1163 (61%) of papers were authored solely by academic authors, 122 (6%) of papers were authored solely by authors with industry affiliations, and 613 (32%) had authors from both academia and industry – see Table 1.

[1] In November 2020 the website contained 1899 papers. One paper was later taken down from the site and has therefore been removed from our dataset, which now includes 1898 papers.

Table 1: Academia versus industry

Affiliation type	Number of papers	Percentage
Academia	1163	61%
Industry	122	6%
Mixed	613	32%

3.1.2 Most represented affiliations. In Table 2 we show all institutions associated with at least 30 accepted papers. The most common affiliations were Google (240 papers, 13%), Stanford (141, 7%), Microsoft (98, 5%), MIT (92, 5%) and UC Berkeley (79, 4%). Together, over 28% of papers have at least one author affiliated with these five institutions. Over 16% of papers have an Alphabet affiliation (such as Google or DeepMind).

3.1.3 Location. We grouped affiliation countries into continents – see Table 3. We found that around two-thirds of papers had at least one North American affiliation (1279 papers, 67%), almost a third had a European affiliation (557, 29%) and likewise for Asian affiliations (551, 27%). There were only 58 papers with an affiliation from Oceania (3%), 10 for South America (0.5%) and only three from Africa (0.2%). Note that the percentages do not sum to 100% since papers can have authors from more than one region. US affiliations make up the majority of those from North America – 65% of all papers have a US affiliation.

A heatmap showing the number of papers with at least one author affiliation associated with each country is shown alongside Table 3. The ten countries with the most associated papers were the US (1229 papers, 65%), China (295, 16%), the UK (222, 12%), Canada (110, 6%), France (106, 6%), Switzerland (94, 5%), Germany (88, 5%), South Korea (59, 3%), Australia (58, 3%) and Israel (55, 3%).

3.1.4 Subject area. There were 162 granular primary subject areas represented, but these are grouped into ten main categories, such as *Deep Learning* or *Theory*. Table 4 shows the number of papers for each category. We note that while there are a range of subject areas represented, a large proportion of papers have primary subject areas indicating work towards the theoretical end of the spectrum, such as *Theory* (195 papers, 10%), *Optimization* (125, 7%), and *Algorithms* (504, 27%). Those in method based subject areas such as *Deep Learning*, *Reinforcement Learning* and *Probabilistic Methods* likely span the theory-application spectrum, including many that focus on general purpose techniques that may be applied to a range of applications.

3.2 Engagement with broader impact statements

3.2.1 Opt-outs. While it was mandatory to include a broader impact section, the NeurIPS template stated that "If authors believe this is not applicable to them, authors can simply state this" [37]. In the FAQs authors were told that "if your work is very theoretical or is general enough that there is no particular application foreseen, then you are free to write that a Broader Impact discussion is not applicable" [19]. Here we investigate how many authors chose to "opt out" in this way.

We made the assumption that any statement of more than 60 words constituted an attempt to include a statement, and labelled these as opt-in. We manually labelled all statements of 60 words or less as opt-out or opt-in depending on their content; there were 361 such statements. While many statements were clearly an opt-out, there were many ambiguous examples. For example, some authors imply that an impact statement is not applicable, but go on to mention real world applications (e.g. [20]), or detailed descriptions of the impact of their work on the field (e.g. [31]). Other ambiguous examples include stating that they do not anticipate any impacts or ethical aspects that are not "well understood by now" [33], or mentioning that their method "brings risk", without elaborating

Table 2: Most represented affiliations - https://public.flourish.studio/visualisation/6152118

Institution	Number of papers	Percentage
Google*	240	13%
Stanford University	141	7%
Microsoft	98	5%
MIT	92	5%
UC Berkeley	79	4%
DeepMind	76	4%
Tsinghua University	74	4%
Princeton University	63	3%
Columbia University	52	3%
ETH Zürich	51	3%
Harvard University	43	2%
Cornell University	41	2%
Facebook	38	2%
Peking University	38	2%
UCLA	38	2%
IBM	38	2%
EPFL	30	2%
UIUC	30	2%

including "Google DeepMind"

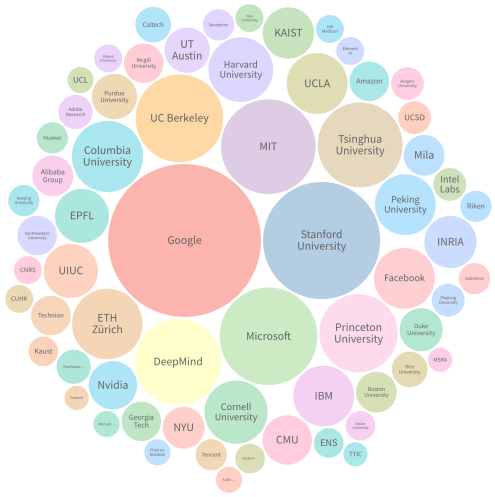

Table 3: Regions - Number and percentage of papers containing at least one affiliation from each region and number of papers associated with each country - https://public.flourish.studio/visualisation/6084514

Region	Number of papers	Percentage
North America	1279	67%
Europe	557	29%
Asia	551	27%
Oceania	58	3%
South America	10	0.5%
Africa	3	0.2%

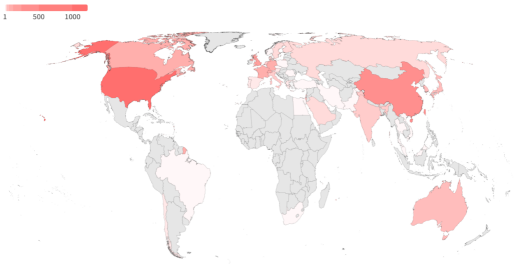

Table 4: Primary subject area

Primary subject area	Number of papers	Percentage
Algorithms	504	27%
Deep Learning	325	17%
Applications	283	15%
Theory	195	10%
Reinforcement learning and planning	183	10%
Optimization	125	7%
Probabilistic methods	125	7%
Social aspects of machine learning	77	4%
Neuroscience and cognitive science	69	4%
Data, challenges, implementations, and software	12	1%

[41]. We labelled 74 statements as ambiguous; 45 of which we decided on balance to include as opt-outs.

In total, we labelled 185 statements (10%) as opt-outs – see Table 5.

Table 5: Opt-outs

	Number of papers	Percentage
Opt-out	185	10%
Opt-in	1713	90%

We found small differences by affiliation type: opt-outs for *Academic*, *Industry*, and *Mixed* constituted 11%, 7% and 9%, though the figure for *Industry* suffers from small sample size (only eight statements in the *Industry* category were labelled as opt-out). See Figure 1. Results by region for *North America*, *Europe* and *Asia* were similar to the average: 9%, 12% and 11% respectively.

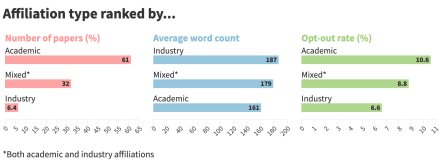

Figure 1: Affiliation type - https://public.flourish.studio/visualisation/6478786

We found more pronounced differences in opt-out rates when disaggregated by subject area. Perhaps unsurprisingly, we found that theoretical topics, such as *Theory* and *Optimization* had the highest opt-out rates (25% and 24%), and *Applications* and *Social Aspects of ML* had the lowest (2% and 1%). See Figure 2. It is worth noting that even for *Theory*, 75% still opted in to at least some degree, despite authors of very theoretical work being told they may simply write that a statement is not applicable.

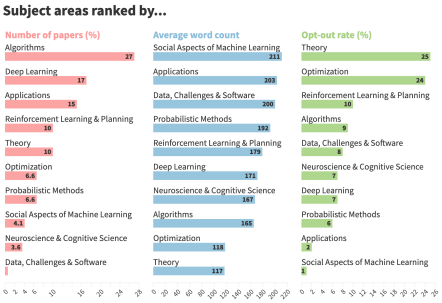

Figure 2: Primary subject area - https://public.flourish.studio/visualisation/6915390

3.2.2 Statement length. The mean statement length was 169 words and 7.3 sentences, i.e. approximately one short paragraph – see Table 6. This is not dissimilar to the example provided by NeurIPS in the FAQs, namely a paper with a 191 word ethics statement [19]. If we look at only those papers labelled as opt-in, the mean length rises to 184 words and 7.9 sentences.

Table 6: Average statement length

	Mean word count	Mean sentence count
All papers	169	7.3
All opt-in papers	184	7.9

While the mean length was 169 words, there was large variation between statements. The shortest statements were two words, namely "Not applicable" (e.g. [54]), and the longest statement was 4337 words and 150 sentences long [12]. The median length was 138 words. See Figure 3 for the entire distribution.

Again, we found small differences between affiliation type, see Figure 1, and location. For example, of the 795 papers whose authors all had US affiliations, and 97 papers with Chinese affiliations, the average word lengths were 182 words and 135 words respectively.

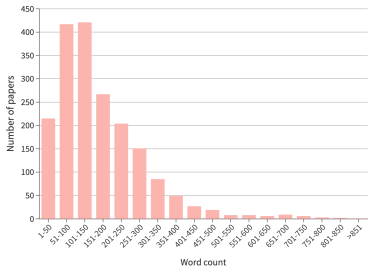

Figure 3: Length of impact statements - https://public.flourish.studio/visualisation/5979712

Similarly to opt-out rates, we found differences between subject areas. *Theory* and *Optimization* had the shortest average statement length (117 and 118 words respectively), and *Social Aspects* and *Applications* had the longest (211 and 203 words) – see Figures 2 and 4.

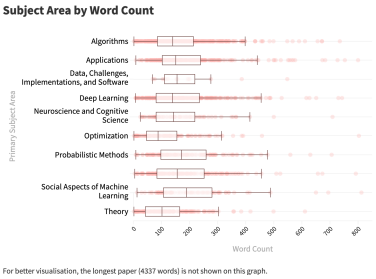

Figure 4: Length distribution by subject area - https://public.flourish.studio/visualisation/6084366

Drilling down into the *Applications* category, we can also see differences between sub-areas. In Table 7, we show mean word counts for all *Application* sub-areas with at least 10 accepted papers. For example, we find that *Health* and *Computational Biology* are well above average (265 and 255 words respectively). Perhaps surprisingly, *Computer Vision* and *Natural Language Processing* (NLP) differ greatly (166 and 340 words respectively). We note that the longest statement (4337 words) is found in an NLP paper, namely OpenAI's GPT3 paper *Language models are few-shot learners* [12]. The average word count for NLP statements excluding this paper is 223 words – around a third longer than those from *Computer Vision*.

3.3 Themes

For a thorough treatment of the themes covered in a large sample of impact statements, see Nanayakkara et al. [34]. To get a flavour of some of the topics discussed in all statements, and whether these differed by affiliation type and subject area, we enumerated the highest frequency words across all impact statements, with stopwords (such as "the" and "and") removed.

The top 30 words can be found in Table 8. As illustrated by the top five words (*learning*, *work*, *data*, *models*, *applications*), the list contains a high proportion of "technical" terms. In Table 9, we

Table 7: Application sub-areas

Primary subject area	Number of papers	Mean word count
Natural Language Processing	35	340
Health	12	265
Computational Biology and Bioinformatics	12	255
Network Analysis	11	236
Time Series Analysis	10	219
Applications [No sub-topic listed]	15	170
Computer Vision	96	166

Table 8: Most common words

Word	Freq	Word	Freq
learning	2516	method	816
work	2257	methods	789
data	1941	paper	786
models	1393	neural	649
applications	1387	many	647
model	1077	deep	635
research	1022	new	611
machine	945	networks	588
systems	886	theoretical	588
used	876	use	565
will	875	proposed	555
impact	872	one	553
algorithms	845	eg	524
potential	840	tasks	510
training	827	societal	497

therefore give a list of the top 30 words that we deemed to be related to societal impacts. We show in bold those terms related to particular classes of societal considerations.

This list gives an indication of some of the most common societal considerations considered, namely privacy (*privacy* occurred 403 times), fairness (*fairness*, *biases*, *bias* occurred 243, 334, 307 times respectively), robustness (*robust*, *robustness*, *adversarial* occurred 309, 260, 296 times), and safety (*safety* occurred 202 times). These themes were fairly consistent across locations and affiliation types. We found the top words associated with application settings were *medical* (238 occurrences), *robots* (123), *science* (123), *malicious* (113), *scientific* (106), *healthcare* (104), *health* (96), *robots* (94), *decision-making* (79), *surveillance* (79) and *industry* (73).

3.4 Valence

To get an indication of whether authors described both positive and negative impacts to the same extent, we created a list of synonyms for *positive* using Wordnet's synset (semantic) relations [49]. We took the number of occurrences of positive synonyms that were not negated as a measure of explicitly positive sentiment. We repeated this for synonyms of *negative*, *strength* and *limitation* – see Table 10. This method has many limitations, including that it does not account for implicit positive impacts. For example, if an author states their work could be applied to healthcare, they may intend for this to be viewed as a positive impact, though no synonym for positive is used. The counts should therefore be taken to be a crude proxy.

Table 9: Top words relating to societal impacts

Word	Freq	Word	Freq
impact	872	community	272
societal	497	researchers	271
human	415	**robustness**	260
privacy	403	broader	257
future	398	risk	254
biases	334	**fairness**	243
realworld	317	study	241
consequences	315	autonomous	240
robust	309	medical	238
bias	307	real	236
social	296	users	224
adversarial	296	world	219
society	291	present	213
ethical	289	people	206
impacts	282	**safety**	202

From Table 10 we see that the number of occurrences of synonyms for *positive* is higher than *negative* (an average of 4.6 occurrences per statement versus 3.6). This quantitative measure supports the observations about valence given by Nanayakkara et al. [34], who describe that some researchers omit discussions of negative impacts, explicitly state that there are no negative impacts, or only include a brief mention of negative impacts. We note a similar bias towards occurrences of *strength* synonyms, over *limitation* synonyms (1.3 versus 0.6 per statement).

Table 10: Valence

Synonyms	Mean occurrences	% statements
Positive	4.6	90%
Negative	3.6	80%
Strength	1.3	60%
Limitation	0.6	37%

4 DISCUSSION

In our analysis, we have found that there is a concentration of authors along geographic and institutional lines. We find high variation in engagement, as measured by length and opt-out rates. We find evidence that certain established topics are common to impact statements, namely privacy, fairness, robustness and safety. We also find evidence that authors tend to discuss positive impacts to a greater extent than negative impacts. We now reflect on what these tell us about the benefits and challenges of broader impact requirements. We use the framing from Prunkl et al. [50], in which

potential benefits, negative outcomes and challenges were identified based on lessons learnt from related governance mechanisms.

4.1 Evidence of benefits

Building on the EPSRC AREA framework, Prunkl et al. [50] categorise the potential benefits of broader impact statements under the headings of *Anticipation*, *Action*, *Reflection and awareness*, and *Coordination*. We find evidence relating to the latter two categories as follows.

Reflection and awareness. Most authors (90%), including the majority of those whose primary subject area falls under theoretical areas such as *Theory* (75%), took the opportunity to reflect on their work and include an impact statement, rather than stating that one was not applicable (§ 3.2.1). In a survey of NeurIPS authors, Abuhamad and Rheault [1] found that of the respondents who supported the requirement, some found the thought process to be the most valuable aspect, suggesting that the opportunity for reflection is valued by some. The long tail of statement lengths shows that several authors took the unlimited page limit as an opportunity to include a thorough treatment of some of the issues they had identified, raising more detailed awareness of such issues among their readers (§ 3.2.2). For example, in *Language models are few-shot learners* (a winner of the NeurIPS 2020 Best Paper Awards), the impact section includes detailed discussions of potential misuse, fairness, and energy use [12]. This paper contained the longest impact statement, and included authors whose main contribution to the paper was detailed analysis of these ethical considerations. For an additional discussion of encouraging trends found in preprint versions of NeurIPS papers, see Boyarskaya et al. [11].

Coordination. These statements have given an indication of some of the issues which many authors felt applied to their work, such as privacy, fairness, robustness and safety (§ 3.3). This correlates with the themes found by Nanayakkara et al. [34], who identified the following issues as themes amongst their sample: bias (24%), robustness and reliability (21%), privacy (19%), environment (10%), interpretability (10%) and labor (6%). Identifying these common themes provides some information about which issues are widely recognised (or otherwise) within the ML community.

Another potential coordinating function of statements is prompting collaboration with other disciplines or stakeholders. While this is difficult to ascertain from the statements directly, Abuhamad and Rheault [1] found that 16% of survey respondents reached out for outside support. Indeed Brown et al. [12] used a contribution section to highlight individuals who worked on fairness analysis, threat analysis and ethical impacts (including the broader impact section).

4.2 Evidence of negative outcomes

Prunkl et al. [50] list quality deficits, trivialisation of ethics and governance, negative attitudes, false sense of security, unintended signalling, and polarisation of the research community as potential negative outcomes of impact statement requirements. Our analysis finds evidence for the first two of these.

Quality deficits. As previously discussed, we found high variation in engagement, as measured by length (§ 3.2.2) and opt-out rates (§ 3.2.1), as well as a bias towards discussing positive aspects (§ 3.4). The average statement is relatively short, around seven sentences (§ 3.2.2). Abuhamad and Rheault [1] found that the vast majority of authors surveyed spent less than two hours on their statement, and just under half spent less than one hour. Some NeurIPS referees remarked on the low quality of statements on social media [50]. Some authors did not engage with negative impacts at all, and some only focused on impacts to their technical field, rather than societal impacts [34].

Trivialisation of ethics. Relatedly, one might interpret the large number of very short statements as evidence of the trivialisation of the societal impacts of NeurIPS work.

Other negative outcomes, such as negative attitudes and polarisation are difficult to gauge directly from the statements themselves; other means such as attitude surveys and deliberative fora are better suited. There was certainly some degree of backlash expressed on social media, and it is possible that negative attitudes towards the requirement was a contributing factor to changing the process for 2021 (the chairs refer to author feedback as one factor leading to the decision [9]).

4.3 Causes and challenges

Here we discuss several challenges relating to impact statements, and potential causes of the negative outcomes identified. Possible causes for quality deficits include lack of explanation and guidance; the complexity of the task (particularly for foundational and general purpose research); high opportunity costs; and institutional, social, cognitive pressure and biases [50].

Institutional, social and cognitive bias. The original motivation given by the NeurIPS chairs was that it was incumbent on the community "to consider not only the beneficial applications and products enabled by our research, but also potential nefarious uses and the consequences of failure" in response to the "more and more pervasive" impact of the community's work [36]. While it was stated that authors must "take care to discuss both positive and negative outcomes", we find evidence that authors tend to discuss positive impacts to a greater extent, with some authors neglecting negative impacts altogether (§ 3.4). One possible contributing factor is that researchers may be incentivized to focus on minor risks or those that do not threaten their own or organisational interests [50].

Because of the high concentration of papers associated with a handful of affiliations (§ 3.1.2), any institutional bias from the most prolific institutions is likely to have a significant effect. Many academic researchers might be affected by industry incentives, not least because of the large number of papers that are collaborations between authors with academic and industry affiliations (§ 3.1.1); see also [23].

The concentration of authors along geographic lines (§ 3.1.3) is also likely to influence how authors view the impacts of their work. We observed different levels of engagement along geographic lines, for example between the US and China (§ 3.1.3), which could be a result of differing attitudes towards ethics and the societal

impact of ML [51]. To date, AI ethics in China has been more government driven than company driven (for example, the government recently released new high-level AI ethics guidelines [52]), whereas American Big Tech companies tend to follow their own codes of ethics in absence of specific government guidelines. This could mean that longer statements from US companies are a result of their companies efforts in this space (which could be in good faith, or could suffer from ethics washing). Language could of course be another contributing factor. Since NeurIPS papers must be written in English we might expect typically longer discussions from those for whom English presents a lower language barrier, which may be more prevalent among US affiliated authors than those affiliated with institutions in China.

Lack of best practice, guidance and explanation of purpose. Although some more detailed unofficial guidance was made available, such as Ashurst et al. [3] and Hecht [24], the *official* guidance given was very limited. A potential cause of the predominance of positive over negative impacts discussed is that authors confused the requirement with other broader impact requirements. For example, the wording surrounding broader impact statements that form part of funding applications often focus on highlighting positive impacts [50]. It is clear that different authors interpreted the aim of the requirement very differently. For example, some only discussed positive societal impacts, some restricted to technical impacts to their field, some focused solely on whether there were new societal impacts introduced by their work [33], some on whether there were immediate impacts (e.g. within the next 6 months [55]) and others focused on other aspects of responsible research such as reproducibility [14].

Conversely, some authors stuck rigidly to the (limited) guidance given. For example, some closely followed the four bullets suggested in the template file [37, 56]. This shows the importance of careful guidance.

It is also possible that lack of communication around the purpose of the requirement (and why it might be valuable) contributed to push back from some researchers. Introducing a new requirement that has associated costs (even a fairly lightweight and flexible requirement, as was arguably the case here) will only be received well if its benefits are understood.

Opportunity costs. Impact statements did not contribute to the page limit, thus removing one potential opportunity cost. This allowed authors to include extended and thorough statements if they so desired. We note that this is no longer the case for 2021 papers: any discussion of negative societal impact counts towards the page limit. While an additional page has been given to account for this (and other new requirements), there is nothing to prohibit authors from using the additional page for other uses, and so any discussion of impacts is in competition with these other uses. There is, however, a downside to the 2020 approach of having a separate section after the page limit: having an impact statement after the main body does separate it from the central work. Some would argue that an integrated approach is more desirable, since societal reflection should be an activity integrated with the research itself.

Additionally, writing a high-quality statement can take time. The high number of very short impact statements, and the short amount of time spent on statements [1], may be at least partly a result of this opportunity cost.

4.4 Conclusion

4.4.1 Summary. The 2020 NeurIPS impact statements present a unique opportunity to investigate the benefits and challenges of this and similar governance mechanisms, and also provide an insight into how ML researchers think about the societal impacts of their work. In order to encourage investigations into how researchers responded to this requirement, we have created a dataset containing the impact statements from all NeurIPS 2020 papers, along with additional information such as affiliation type, location and subject area. We also provide a visualisation tool for exploration and an initial quantitative analysis of the dataset. We investigate the voices represented (§ 3.1), levels of engagement (§ 3.2), themes (§ 3.3) and valence (§ 3.4), and discuss how these reflect the benefits (§ 4.1), negative outcomes (§ 4.2) and challenges (§ 4.3) associated with broader impact statements. We encourage others to investigate the dataset of impact statements, and to continue to further analyse how researchers respond to changing requirements, such as the checklist approach, over the coming years.

4.4.2 What we can learn from 2020 statements. The 2020 requirement provides lessons for self-governance mechanisms more broadly, in particular:

The importance of creating the right incentives. While the requirement was mandatory, there were few incentives to engage deeply with the task, nor any requirement that statements meet any particular standard of quality. While some opportunity costs were accounted for (namely page length), in many ways the process incentivised minimal engagement. For any self-governance mechanism, one must carefully consider how individuals are incentivised to act, and whether this is likely to result in the desired aims.

The importance of clear expectations and guidance. As discussed, there may have been confusion between this requirement and other impact statement requirements which focus on potential benefits. The range of approaches taken by authors show that the aim and expectations were not universally understood. Even within the limited guidance given, it was not clear which elements of responsible research were being targeted. There is therefore a need to disentangle the different elements of responsible research (such as reproducibility, protections for human and data subjects, and anticipation of future downstream consequences), to target mechanisms specifically, and to be clear about the aims. On this occasion, outside guidance did provide some additional help for researchers ([3],[24]); clearer communication of the aims and expectations also better enable external researchers to provide appropriate supporting materials such as these.

The importance of transparency, and constructive deliberation. The wide range of approaches, levels of engagement, and responses to the requirement show a continued lack of consensus around the purpose of mechanisms such as impact statements. Continued deliberation within the community, in consultation with impacted stakeholders and societal experts therefore continues to be of great

importance.

More specifically, what can the 2020 statements tell us about the utility of impact statement requirements? Unfortunately, it is challenging to judge the full potential of impact statement requirements based on one year alone, particularly when little guidance (and warning) was available to researchers. It takes time to build up best practice, good guidance, and effective incentives. We have identified a range of challenges which could be addressed over time to improve outcomes if such a requirement were to persist. However, we have already seen some evidence of the benefits from just one year of the requirement: including providing the opportunity for some researchers to engage thoughtfully, and the opportunity to include extended statements that thoroughly discuss a range of concerns to be addressed. We found that most authors (including those in the most theoretical categories) showed some degree of engagement, promoting reflection and raising awareness of some of the potential risks. Regarding potential negative outcomes, many of these need to be measured through other means, such as attitude surveys. We did find evidence that many statements had limited engagement with the requirement, with some failing to acknowledge any negative consequences. Again, to judge the full potential, this needs to be measured over time. Although monitoring how researchers address societal impacts in their papers may be more challenging under the new checklist requirement (since societal impacts are not required to be in a separate clearly labelled section, but may be discussed throughout the paper), we urge the community to continue to track how researchers choose to do this.

Our hope is that as the ML community continues to grapple with its responsibilities and experiment with different governance mechanisms to encourage responsible research, we will move towards becoming a more mature field with respect to ethical ML. We would like to see societal thinking become a more integrated part of ML research, informing which research is undertaken, and how it is executed. This is of particular importance for research towards the application end of the spectrum, and for research involving data generated by or about people. Until then, we hope the community will (i) continue to experiment with and test governance mechanisms and their associated incentives, (ii) continue to engage in open deliberation both within the community and with outside engagement, and (iii) continue to raise awareness of identified potential harms. While it is encouraging to see that certain classes of harms have become relatively widely recognised (such as privacy and bias), we hope the community will increase its awareness and understanding of a wider range of harms and issues.

ACKNOWLEDGMENTS

For discussion and input we thank: Markus Anderljung, Jeff Ding, Noemi Dreksler, Ben Garfinkel, Matthij Maas, Carina Prunkl and Toby Shevlane. We also thank Earl Ng for the code to extract the impact statements [42], and for providing code review.

Founding/Sources: The authors declare no sources of funding for this 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* (2020).
[2] Affiliation types 2021. https://docs.google.com/spreadsheets/d/1CT3hCvbKxyJeS1FdrZtlK5MTuvWsfXR_
[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] Carolyn Ashurst, Solon Barocas, Rosie Campbell, Deb Raji, and Stuart Russell. 2020. Navigating the Broader Impacts of AI Research. https://ai-broader-impacts-workshop.github.io/
[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[6] Emily Bender. 2019. Is there research that shouldn't be done? Is there research that shouldn't be encouraged? https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6
[7] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).
[8] Michael S Bernstein, Margaret Levi, David Magnus, Betsy Rajala, Debra Satz, and Charla Waeiss. 2021. ESR: Ethics and Society Review of Artificial Intelligence Research. *arXiv preprint arXiv:2106.11521* (2021).
[9] 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
[10] Carolyn Bishoff and Lisa Johnston. 2015. Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University. *Journal of Librarianship and Scholarly Communication* 3, 2 (Sept. 2015), 1231. https://doi.org/10.7710/2162-3309.1231
[11] 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 System* (2020).
[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
[13] Coalition for Critical Technology. 2020. Abolish the #TechToPrisonPipeline. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16
[14] Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, and Pierre-Yves Oudeyer. 2020. Language as a Cognitive Tool to Imagine Goals in Curiosity Driven Exploration. *Advances in Neural Information Processing Systems* 33 (2020).
[15] Emily Denton and Timnit Gebru. 2020. Tutorial on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2020. https://sites.google.com/view/fatecv-tutorial/
[16] EMNLP. 2020. Call For Papers. https://2020.emnlp.org/call-for-papers
[17] EMNLP. 2021. Call For Papers. https://2021.emnlp.org/call-for-papers
[18] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.
[19] NeurIPS FAQ. 2020. NeurIPS 2020 FAQ for Authors. https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28
[20] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. COOT: Cooperative hierarchical transformer for video-text representation learning. *Advances in Neural Information Processing Systems* 33 (2020).
[21] Github repository 2021. https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements
[22] Don Gotterbarn, Amy Bruckman, Catherine Flick, Keith Miller, and Marty J Wolf. 2017. ACM code of ethics: a guide for positive action.
[23] Thilo Hagendorff and Kristof Meding. 2021. Ethical considerations and statistical analysis of industry involvement in machine learning research. *AI & SOCIETY* (2021), 1–11.
[24] Brent Hecht. 2020. Suggestions for Writing NeurIPS 2020 Broader Impacts Statements. https://brenthecht.medium.com/suggestions-for-writing-neurips-2020-broader-impacts-statements-121da1b765bf
[25] 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. https://acm-fca.org/2018/03/29/negativeimpacts
[26] Jeremy Hsu. 2019. Microsoft's AI Research Draws Controversy Over Possible Disinformation Use. https://spectrum.ieee.org/microsofts-ai-research-draws-controversy-over-possible-disinformation-use
[27] Sergey Ivanov. 2020. NeurIPs affiliation locations. https://github.com/nd7141/icml2020/blob/master/university2.csv
[28] Sergey Ivanov. 2020. NeurIPs authors and affiliations. https://github.com/nd7141/icml2020/blob/master/neurips_2020_accepted.txt
[29] Khari Johnson. 2020. NeurIPS requires AI researchers to account for societal impact and financial conflicts of interest. https://venturebeat.com/2020/02/24/neurips-requires-ai-researchers-to-account-for-societal-impact-and-financial-conflicts-of-interest
[30] Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell, and Marc Aurelio Ranzato. 2020. What we learned from NeurIPS 2020 reviewing process. https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f
[31] Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, and Wenxuan Zhong. 2020. Sufficient dimension reduction for classification using principal optimal transport direction. *Advances in Neural Information Processing Systems* 33 (2020).
[32] Jessica Morley, Caio C.V. Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine* 260 (Sept. 2020), 113172. https://doi.org/10.1016/j.socscimed.2020.113172
[33] Vrettos Moulos. 2020. Finite-Time Analysis of Round-Robin Kullback-Leibler Upper Confidence Bounds for Optimal Adaptive Allocation with Multiple Plays and Markovian Rewards. *Advances in Neural Information Processing Systems* 33 (2020).
[34] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2021).
[35] NeurIPS. 2020. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). https://proceedings.neurips.cc/paper/2020
[36] NeurIPS. 2020. Getting Started with NeurIPS 2020. https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28
[37] NeurIPS. 2020. LaTeX template. https://media.neurips.cc/Conferences/NeurIPS2020/Styles/neurips_2020.tex
[38] NeurIPS. 2020. Neurips 2020 Call For Papers. https://nips.cc/Conferences/2020/CallForPapers
[39] NeurIPS. 2021. NeurIPS 2021 Call for Papers. https://neurips.cc/Conferences/2021/CallForPapers
[40] NeurIPS. 2021. NeurIPS 2021 Paper Checklist Guidelines. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist
[41] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems* 33 (2020).
[42] Earl Ng. 2020. Academic pdf scrap. https://github.com/earlng/academic-pdf-scrap
[43] Safiya Umoja Noble. 2018. *Algorithms of oppression.* New York University Press.
[44] Notes on Data 2021. https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements/blob/main/main-dataset/notes-on-data.md
[45] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[46] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.
[47] PAI. 2021. Managing the Risks of AI Research. https://www.partnershiponai.org/responsible-publication-recommendations
[48] Paper subject areas 2021. https://docs.google.com/spreadsheets/d/10HHHgXGAwsoeF6XxX0ZTJUAvixMO2NedcHrPnFm6JfQ
[49] Princeton University. 2010. About WordNet. https://wordnet.princeton.edu/
[50] 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.
[51] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & SOCIETY* 36, 1 (2021), 59–77.
[52] Xinmei Shen. 2021. Chinese AI gets ethical guidelines for the first time, aligning with Beijing's goal of reining in Big Tec. https://www.scmp.com/tech/big-tech/article/3150789/chinese-ai-gets-ethical-guidelines-first-time-aligning-beijings-goal
[53] Visualisations 2021. https://public.flourish.studio/story/961580
[54] Joong-Ho Won. 2020. Proximity Operator of the Matrix Perspective Function and its Applications. *Advances in Neural Information Processing Systems* 33 (2020).
[55] Qiong Wu, Felix Ming Fai Wong, Zhenming Liu, Yanhua Li, and Varun Kanade. 2020. Adaptive Reduced Rank Regression. *Advances in Neural Information Processing Systems* 33 (2020).
[56] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. *Advances in Neural Information Processing Systems* 33 (2020).
[57] Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019.* Profile books.