Ethics consideration sections in natural language processing papers

Luciana Benotti
Universidad Nacional de Córdoba
Via Libre, CONICET, Argentina
luciana.benotti@unc.edu.ar

Patrick Blackburn
Philosophy and Science Studies
IKH, Roskilde University, Denmark
patrickb@ruc.dk

Abstract

In this paper we present the results of a manual classification of all ethical consideration sections for ACL 2021. We also compare how many papers had an ethics consideration section per track and per world region in ACL 2021. We classified papers according to the ethical issues covered (research benefits, potential harms, and vulnerable groups affected) and whether the paper was marked as requiring ethics review by at least one reviewer. Moreover, we discuss recurring obstacles we have observed (highlighting some interesting texts we found along the way) and conclude with three suggestions. We think that this paper may be useful for anyone who needs to write — or review — an ethics section and would like to get an overview of what others have done.

1 Introduction

The first conference of the Association for Computational Linguistics (ACL) to include an ethics advisory committee was the one organized by its North American Chapter (NAACL) in 2021. Since then all ACL conferences have had one, including all chapter conferences, and the largest conferences ACL-IJCNLP and EMNLP. The ACL 2021 webpage call for papers says the following.[1]

> *Authors will be allowed extra space after the 8th page for a broader impact statement or other discussion of ethics. Note that though the ethical consideration paragraph is not mandatory, authors of papers working with sensitive data or on sensitive tasks that do not sufficiently discuss these issues may receive a conditional acceptance recommendation.*

This paper examines the current state of the NLP research community's discussion of the ethical impact of its work, as reflected in its publications. We first pose five empirical questions, and then explore possible obstacles to the discussion. The paper is organized as follows. Section 2 briefly reviews related work, giving references relevant to several areas of NLP. In Section 3 we explore the following five questions regarding ethical consideration sections in ACL 2021 papers:

- What percentage of papers include an ethical considerations section (ECS)?
- Are there some tracks that stand out, either positively or negatively?
- What types of ethical questions are addressed in ECSs, and in what proportion?
- Are the papers with ECSs that went through ethics review different from those that did not?
- Are there differences between countries with respect to the ECSs?

In Section 4 we consider obstacles that may impede the discussion of ethical issues in research papers, and how they might be overcome. In Section 5 we propose three concrete suggestions that can be useful in the future for authors, reviewers and ethics chairs. After the conclusions in Section 6 we discuss the ethical considerations and the limitations of this work.

2 Related previous work

Since the First ACL Workshop on Ethics in Natural Language Processing in 2017, the ACL has initiated regular discussions of various aspects of ethics and the social impacts of NLP. Ethical guidelines have been proposed for broad research areas such as Machine Translation (Haroutunian, 2022) and Natural Language Generation (Smiley et al., 2017) and for more specialized topics, for example, Text Simplification (Gooding, 2022). Resources exist for NLP in general (Leidner and Plachouras, 2017), for AI in the community (Mohammad, 2022), and for considerations relevant when designing shared tasks or benchmarks (Parra Escartín et al., 2017).

There are many references relevant to writing a good ECS. We have no space to discuss them,

[1] Source https://2021.aclweb.org/calls/papers/

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4509–4516
December 7-11, 2022 ©2022 Association for Computational Linguistics

but we note the following references relevant to responsible data usage (Drugan and Babych, 2010; Couillault et al., 2014; Mieskes, 2017; Bender and Friedman, 2018; Kann et al., 2019; Rogers et al., 2021; Gebru et al., 2021), crowdsourcing (Snyder, 2010; Bederson and Quinn, 2011; Fort et al., 2011; Callison-Burch, 2014; Fort et al., 2014; Hara et al., 2018; Toxtli et al., 2021), biases (Blodgett et al., 2020), language diversity (Tatman, 2017; Jurgens et al., 2017; Zmigrod et al., 2019; Tan et al., 2020; Koenecke et al., 2020; Bird, 2020), rigorous and meaningful evaluation (Caglayan et al., 2020; Ethayarajh and Jurafsky, 2020; Antoniak and Mimno, 2021; Tan et al., 2021), environmental impact (Strubell et al., 2019; Zhou et al., 2020; Henderson et al., 2020; Schwartz et al., 2020; Bannour et al., 2021; Przybyła and Shardlow, 2022), and human harms and values (Winner, 1980; Hovy and Spruit, 2016; Leidner and Plachouras, 2017).

3 Methodology and empirical results

In this section we first describe how we answered the questions posed in the introduction, and then present and discuss our findings.

3.1 Methodology

We begin by describing the aspects of ethics reviewing we annotated automatically, and those we annotated manually. We took 572 papers accepted as long papers at the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021). We first automatically searched for those containing such section titles as *Ethical Considerations*, *Ethical Concerns*, *Social Impact* and *Broader impact statement* and variations. We found 90 such papers, that is, 15.7% of the total number published. By way of comparison, the following year (while this paper was being written), 604 long papers were published at ACL 2022, of which 141 had an ethics consideration section, thus 23.3% of published papers had an ECS.

In their final conference report, the ACL 2021 ethics advisory committee writes:[2]

> *After conducting our review, 28 papers were given Conditional Accept decisions by our committee, meaning they were asked to submit changes prior to final acceptance. No papers were rejected by our committee.*

As shown in column R on Table 1, 48 papers went through an ethics review. Ethics reviewers could decide whether a paper was to be: (a) accepted as is, (b) accepted with minor changes, or (c) conditionally accepted. The above quote states that 28 papers were conditionally accepted in ACL 2021. This means that the acceptance of the majority of papers that went through an ethics review was conditional on the paper being modified, and that reviewers verified that suitable changes were made before publication. This might be a reason why the ECSs of the papers that went through the review contain more discussion of harms and vulnerable groups (we return to this point in Section 3.2).

We read the 90 ethical considerations sections of the ACL 2021 papers and annotated whether the section mentioned an answer to the following questions (any disagreements between the two annotators were discussed and agreed upon). We did not evaluate the quality of the answers, merely noted whether the section mentioned the topic.

Benefits: If the technology functions as intended, who **benefits**?

Harms: If the technology functions as intended, or if it fails, who might be **harmed**, and how?

Vulnerabilities: Are any of the possible harms you've identified likely to fall disproportionately on populations that already experience marginalization or are otherwise **vulnerable**?

The three questions we pose represent all topics mentioned by the 6 points on the EMNLP2022's Ethics FAQ. Point 1 in the FAQ refers to benefits and points 2 to 6 refer to particular kinds of harms; namely harms related to failure, bias, misuse, data collection, and vulnerable groups, respectively. Thus our harms category covers items 2-5 in the FAQ, and we separately annotate harms related to vulnerable groups as vulnerabilities. This is because vulnerabilities are a particular kind of harm, one that seems better defined than the others. By way of contrast, what constitutes misuse, for example, is less clear.

3.2 Empirical results

As we said earlier, 15.7% of the papers published at ACL 2021 contained an ECS; of these, 53% were evaluated by the conference ethics review committee as shown in Table 1 (column %R). Our classification of all papers with an ECS shows that 74% of them describe the research benefits, 52% describe

[2] The full report: https://tinyurl.com/yswmnu54

	#	S	B	%B	H	%H	V	%V	R	%R
China	231	21	18	86	10	48	3	14	15	71
USA	168	42	28	67	21	50	9	21	24	57
Europe	104	17	13	76	13	76	6	35	7	41
Others	69	10	8	80	3	30	3	30	2	20
Total	572	90	67	74	47	52	21	23	48	53

Table 1: Distribution of papers whose first author is affiliated in China, the United States, Europe or elsewhere, with information regarding the ECSs they contain. The first column is the number of total papers. S is the number of papers with an ECS, B those papers whose ECS describe research benefits, H for those that describe harms, V for those that describe harms to vulnerable groups, and R those papers that were marked by at least one reviewer as requiring an ethics review. %H, %B, %V, %R is calculated as the percentage of H, B, V, R respectively with respect to S.

potential harms, and 23% describe harms that particularly affect vulnerable groups. Moreover, 20% of the papers do not address any of the questions we just listed; most of these consist of a text stating that the work has no risks, that crowdworkers were fairly compensated, or that the research was approved by an IRB. We call such ECSs ‘disclaimers’, and discuss them in Section 4.

Table 1 also classifies papers by the country of residence of the first author; of course, another author may have had a stronger influence on whether to include an ECS, and on what its contents would be, than the first author. Measured this way, the table shows that China is the country with most accepted papers (231 papers is 40% of the submissions) and also the country with the highest proportion of papers (.71 in column %R vs .41 in Europe[3]) that have an ECS that went through an ethics review. The publication of papers that go through an ethics review is usually conditional on them adding an ECS that discusses the issues raised by the committee. For the USA (at 30%, the second most highly ranked country in terms of accepted papers) this proportion goes down to .57. This means that more papers from the US (than from China) that were not required to have an ECS, have one. However, when we look at the content of the ECSs, the US and China do not look so different. The proportion of ECSs that discuss harms (column %H) is close: .50 for USA and .48 for China.[4] ECSs from both countries discuss benefits in higher proportion than harms: .86 vs .67 in column %B for China and the USA respectively. We also observed that a considerable proportion of ECSs from the USA just contained a disclaimer. Of all the papers published, 18% come from Europe and 16% of them contain an ECS. European ECSs seem rather different (though the numbers are smaller): the same proportion of papers in Europe discusses both benefits and harms (.76 in columns %B and %H). Europe also has the largest proportion of papers that discuss harms to vulnerable groups (.35 in column %V).

Over 1/3 of the ethical consideration sections are quite short, with less than 19 lines each as shown in Figure 1. The average (mean) length is 33 lines. The average length of the ethical consideration section evaluated by the ethics review committee is 46 lines, while the average length of unevaluated sections is 26.

Published papers that were marked (by at least one scientific reviewer) as needing to be reviewed by an ethics committee are distributed across many different tracks; see Figure 2. The track with the highest percentage was *Resources and Evaluation* with 14.6% followed by *Computational Social Science and Cultural Analytics* and *Information Extraction* with 10.4%; the usual assumption that *NLP Applications* is a track with more potential to generate ethical conflict does not seem to hold.[5]

As we mentioned earlier, the ethical consideration section of those papers marked as needing to be reviewed by an ethics committee tended to be longer (46 versus 26 lines) and discussed more harms. We doubt that forcing authors to add an ethics consideration section is useful. However,

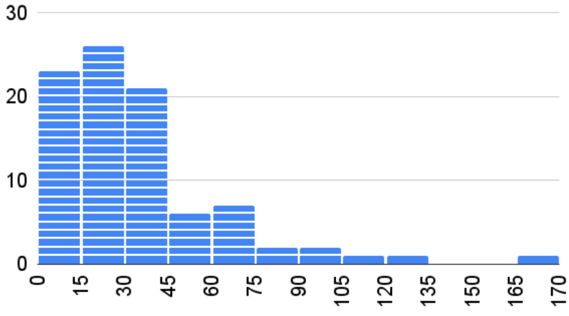

Figure 1: Histogram of the length of the ethical considerations sections in number of lines of the 90 papers that include such section in ACL 2021.

[3] For a two proportion Z-test, the difference is statistically significant, the value of p is .03144

[4] For a two proportion Z-test, the difference is not statistically significant.

[5] https://2022.emnlp.org/ethics/faq/

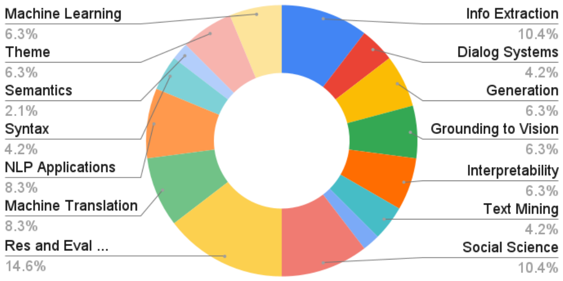

Figure 2: Papers by track that were marked as *it needs to be reviewed by an ethics committee* by at least one scientific reviewer of the paper.

ethics reviewing may be helping NLP researchers think deeper about their ECSs. We empirically compared those papers that were evaluated with respect to the three questions concerning benefits, harms and vulnerabilities. There was not much difference in the discussion of benefits, with .78 for papers that got a review versus .72 for those that did not (researchers are already motivated to do discuss benefits). But reviewed papers are more likely to describe harms (.69 vs .43) and how they affect vulnerable groups of people (.38 vs .12).

4 Overcoming obstacles

In this section we reflect on the obstacles facing the discussion of ethical issues by the NLP community. Perhaps the most basic question to pose here is whether ethical issues can usefully be discussed in ECSs in research papers. In our experience, there are many people in the NLP community who are not keen on the idea. Sometimes this seems to come from worries about doing it well: *I'm a computer scientist - what do I know about ethics?* or perhaps *My English is ok for writing technical papers, but no way could I write about benefits and harms*. But we have also heard frustrations about the very idea of writing ECSs: *they are a waste of time, all that has to be said already has been said*, that there simply is nothing serious to say here (*When I asked my supervisor about this, he laughed*). In short: it is nothing but a waste of valuable research time.

But many authors *do* have interesting things to say. The following is from an ECS on the track *Interpretability and Analysis of Models* for NLP (Zhang et al., 2021) reflecting on vulnerable groups (see boldface) and containing critical evaluations of harms and benefits:

> *There are several ethical reasons to study LMs with limited pretraining data. Training massive LMs like RoBERTa from scratch comes with non-trivial environmental costs (Strubell et al., 2019), and they are expensive to train,* ***limiting contributions to pretraining research from scientists in lower-resource contexts****. By evaluating LMs with limited pretraining, we demonstrate that smaller LMs match massive ones in performance in many respects. We also identify a clear gap in our knowledge regarding why extensive pretraining is effective. Answering this question could lead to more efficient pretraining and ultimately reduce environmental costs and make NLP more accessible.*

Here is an example of an ECS in the *Ethics in NLP* track (Blodgett et al., 2021) reflecting on potential harms when it is hard to see them.

> *Work concerning the fairness, transparency, or ethics of computational systems is often taken to be inherently beneficial with little to no potential for harm, and thus often (paradoxically) fails to examine its limitations or possible unintended negative consequences (Boyarskaya et al., 2020). And yet, our work is not without risks either; we risk discouraging the type of work we actually want to encourage, and dissuading practitioners from using existing benchmarks to test their models.*

The next example shows that it is possible to discuss the benefits of the research without overselling it in the *Dialogue and Interactive System Track* track (Liu et al., 2021).

> *Considerable additional work is needed to determine what are appropriate levels of support for systems to provide or that can be expected from systems, but our work provides a cautious, yet concrete, step towards developing systems capable of reasonably modest levels of support. The corpus we construct can also provide examples to enable future work that probes the ethical extent to which systems can or should provide support. In addition to these broader ethical considerations, we have sought to ethically conduct this study, including by transparently communicating with crowdworkers about data use and study intent, compensating workers at a reasonable hourly wage, and obtaining study approval from the Institutional Review Board.*

Over the years we have also heard many statements to the effect that *my work is completely theoretical — it raises no ethical issues at all.* Thus we also found the next ECS insightful. It is from a paper that links theoretical work on transfer learning with risks for language diversity (Ahmad et al., 2021), and the authors made the following point:

> *We discuss the limitations and challenges in utilizing universal parsers to benefit the pretrained models. Among the negative aspects of our work is the lack of explanation why some languages get more benefits over others due to universal syntax knowledge incorporation.*

However, as mentioned earlier, another type of ECS caught our attention: what we called 'disclaimers'. These sometimes draw on factual statements—for example, noting that the project leading to the research had been classified as no risk—but sometimes rely on more subjective judgements, for example, that the work raises no ethical issues not already raised by established NLP applications.

We would immediately like to emphasize that there is nothing wrong with disclaimers. The fact that a review board has classified research as (say) low risk may be worth noting—though to say *only that* seems unlikely to raise many interesting ethical issues. What is potentially concerning, however, is that this type of ECS could become a template, perhaps something to teach new PhD students (*You write the abstract like so, the acknowledgments like so, and the ECS like so—got it?*). All well and good—but if folding ECSs into the structure of a research paper becomes a mechanical exercise, this could tame their potential.

This leads to tough questions. What are ECSs meant to achieve? To encourage self-reflection among researchers? To signal compliance with institutionalized norms of best practice? How can ECSs best be used to reach these goals? Whatever the answers, we think it will be enlightening to track the different types of ECSs (as we did in this paper) in future ACL events.

5 Three concrete suggestions

We conclude here with three concrete suggestions. First (for organizers) we suggest that they give authors the option of asking their paper to be reviewed by an ethics committee. One of our findings is that ethics review can be an educative process in which authors may gain a better understanding of the impact of their work. Second (again for organizers) all the ACL 2021 papers are available in the anthology with their videos, but almost none mention ethical considerations; we propose that subsequent conferences offer (optional) extra video time for ethical commentary. Thirdly, in October 2021 the ACL established a stable Ethics Committee to provide consistency between conferences. One of its goals is to build the ACL Ethics Union Bibliography. This is a public list moderated by the current ACL Ethics Committee; you issue a pull request against the repository to have your suggestions discussed before they are approved for integration with the list.[6] Any contributor with an approved suggestion can add their name to the list of contributors to the bibliography. Contributions can be assigned a topic tag in the areas we discussed in Section 2. We suggest that all ACL members check it out!

6 Conclusions

We believe our paper may be useful for authors writing an ethics section — and perhaps even for reviewers of such sections — who would like to get an overview of what others have done, including what might be addressed (benefits, harms, vulnerable groups).

This paper provides a dive into the contents of ethical consideration sections (ECS) in ACL proceedings papers from last year. Specifically, the paper compares the percentages of papers containing ECS from different countries and the contents of these ECS. We also discuss the types of obstacles facing ECS writers, and provide concrete suggestions for authors, reviewers, and organizers.

We wrote this paper because we think it is important to inspire the community to think about the social impacts of their work. We value diverse perspectives — so we do not supply cut-and-dried solutions to the issues we raise. Rather, we have tried to make ethical discussions more meaningful by drawing attention to what has already been done in ECSs. The statistics we reported are simple and basic, and doubtless more detailed analyses will be made as more data becomes available. But our main goal here was to raise these issues clearly and directly — and as soon as possible — to the NLP community.

[6] See https://github.com/acl-org/ethics-reading-list.

7 Ethical considerations

Ethics reviewing in NLP was first implemented in June 2021 for NAACL 2021. In this paper we addressed the following questions about papers published at ACL 2021 in August 2021. 1) What percentage of papers include an ethical considerations section (ECS)? 2) Are there some tracks that stand out, either positively or negatively? 3) What types of ethical questions are addressed in ECSs, and in what proportion? 4) Are the papers with ECSs that went through ethics review different from those that did not? 5) Are there differences between countries with respect to the ECSs?. We also describe common obstacles and arguments regarding ECSs and illustrate papers that have overcome them insightfully. Potential harms of our paper include over-generalizations of the empirical results we show here so we want to make our limitations explicit and we do so in the next section. One of our reviewers pointed out that a potential harm is that this paper raises opinions about the Ethics Consideration Section, which is to some extent sensitive, and that may affect the point of view of other authors toward ECS. To this we can only say: trying to raise awareness and stimulate open discussion of ECSs in the NLP community seems better than leaving them unexamined..

We believe this paper might benefit NLP researchers who are authors, reviewers or conference organizers in different ways. Authors might find in this paper tools to come up with better ECSs. Ethics reviewers might see the impact of their effort. And organizers could have a glimpse at questions addressed and not addressed by ECSs. Our goal is to contribute to the ongoing debate on what is the current situation of the broader societal impact discussions of NLP research.

8 Limitations

This paper is merely a static picture of the state of affairs for ACL 2021 which may very well change in the future. In particular, our analysis per country is limited in that most countries in the world are not represented. In this paper we use the country self-reported by the first author of the paper. There were only two annotators for the contents of ECS. The reported statistics are simple; we only report absolute numbers and percentages. The number of annotated papers is small. It is therefore difficult to know to what extent this is a momentary snapshot and which trends will persist over time. It would be good for future work to include the analysis of later conferences.

Acknowledgements

We want to thank the reviewers for their (often very detailed) comments and suggestions. We would also like to thank the organizers of the Widening NLP (WiNLP) Workshop for financially supporting the in-person presentation of this paper. In particular, we would like to thank Xanda Schofield. We are also grateful to the ACL Ethics Committee for discussion and inspiration.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

Benjamin B. Bederson and Alexander J. Quinn. 2011. Web workers unite! addressing challenges of online laborers. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 97–106, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chris Callison-Burch. 2014. Crowd-workers: Aggregating information across turkers to help them find higher paying work. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2(1):8–9.

Alain Couillault, Karën Fort, Gilles Adda, and Hugues de Mazancourt. 2014. Evaluating corpora documentation with regards to the ethics and big data charter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4225–4229, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jo Drugan and Bogdan Babych. 2010. Shared resources, shared values? ethical implications of sharing translation resources. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 3–10, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Karën Fort, Gilles Adda, Benoît Sagot, Joseph Mariani, and Alain Couillault. 2014. Crowdsourcing for language resource development: Criticisms about amazon mechanical turk overpowering use. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 303–314, Cham. Springer International Publishing.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54, Dublin, Ireland. Association for Computational Linguistics.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad

Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Saif Mohammad. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, Dublin, Ireland. Association for Computational Linguistics.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Piotr Przybyła and Matthew Shardlow. 2022. Using NLP to quantify the environmental cost and diversity benefits of in-person NLP conferences. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3853–3863, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics.

Jeremy Snyder. 2010. Exploitation and sweatshop labor: Perspectives and issues. *Business Ethics Quarterly*, 20(2):187–213.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. *ACM Human Computer Interaction*, 5:1–26.

Langdon Winner. 1980. Do artifacts have politics? *Daedalus*, 109(1):121–136.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Sharon Zhou, Alexandra Luccioni, Gautier Cosne, Michael S Bernstein, and Yoshua Bengio. 2020. Establishing an evaluation metric to quantify climate change image realism. *Machine Learning: Science and Technology*, 1(2):025005.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.