On the Utility of Lay Summaries and AI Safety Disclosures: Toward Robust, Open Research Oversight

Allen Schmaltz
Harvard University
schmaltz@fas.harvard.edu

Abstract

In this position paper, we propose that the community consider encouraging researchers to include two riders, a "Lay Summary" and an "AI Safety Disclosure", as part of future NLP papers published in ACL forums that present user-facing systems. The goal is to encourage researchers–via a relatively non-intrusive mechanism–to consider the societal implications of technologies carrying (un)known and/or (un)knowable long-term risks, to highlight failure cases, and to provide a mechanism by which the general public (and scientists in other disciplines) can more readily engage in the discussion in an informed manner.

This simple proposal requires minimal additional up-front costs for researchers; the lay summary, at least, has significant precedence in the medical literature and other areas of science; and the proposal is aimed to supplement, rather than replace, existing approaches for encouraging researchers to consider the ethical implications of their work, such as those of the Collaborative Institutional Training Initiative (CITI) Program and institutional review boards (IRBs).

1 Introduction

Recent research advances in natural language processing have the potential to translate into real-world products and applications. As with the broader field of artificial intelligence (AI), more generally, there is not a broad consensus on whether the long-term social impact of such advances will be positive or negative–and to whom any future negative impacts will be most acutely dealt. However, there is perhaps consensus that it is useful for researchers to at least consider the potential societal impacts of their work. The concern is not entirely speculative, as user-facing applications of NLP today in areas such as education, for example, have the potential to have large proportions of users who are minors and/or members of at-risk groups, with the output of such systems used in high-stakes educational assessment.

To encourage NLP researchers to consider the societal impacts of their work and to involve the general public in the discussion, we propose that the community consider encouraging authors–on a voluntary basis field-tested in a workshop setting–to include two riders for papers describing user-facing systems or methods. One, a "Lay Summary", which has precedence in journals in other scientific fields, is a short summary aimed at a non-specialist audience designed to reduce misinformation and engage the public. The second, an "AI Safety Disclosure", is a brief overview of potential failure scenarios of which real-world implementations, downstream applications, and future research should be aware.

We surmise that the utility of these riders will be particularly high for NLP papers for which the proposed approaches or methods are aimed at eventually building user-facing systems (e.g., for machine translation, grammar correction, or summarization), but for which the actual research did not directly involve human subjects and thus (rightly so), fall outside the purview of traditional mechanisms such as institutional review boards.

2 Proposal

We propose that NLP articles presenting user-facing systems or methods include two riders, a "Lay Summary" and an "AI Safety Disclosure", as explained further below. By user-facing system or method, we refer to tasks in which the end consumer of the output is a human for performing a real-world task. This would include papers on tasks such as machine translation and summarization, even if the research itself did not involve

Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing, pages 1–6
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

human subjects. It would exclude papers of a more theoretic nature, or for which the end goal is not user-facing output. For example, this criteria might reasonably exclude a paper introducing a new approach for dependency parsing (McDonald et al., 2005) or an empirical comparison of language models (Chen and Goodman, 1996), but it would include papers using dependency parsing or language models as part of a downstream task, such as machine translation. As with other aspects of this proposal, we leave it to the discretion of authors as to whether their paper meets this criteria, and the community may desire to restrict or expand the determination of which papers should include these riders (see Section 4).

Lay Summary The idea of including a summary of an article that is accessible to a general audience is a well-established concept, implemented in existing journals in a variety of scientific fields. Such a summary can assist science journalists and inform discussions in public forums. To a lesser extent, such summaries can also be useful for researchers in other branches of science and engineering.

The journal *Autism Research*, for example, requires a lay summary of "2-3 sentences (60-80 words; 300-500 characters including spaces) included at the end of the Abstract that summarizes the impact/importance/relevance/key findings of the study"[1]. In a similar vein, the *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* requires authors to provide a "120-word-maximum statement about the significance of their research paper written at a level understandable to an undergraduate-educated scientist outside their field of specialty. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership."[2] Shailes (2017) collected a list of 50 journals across the sciences that provide such summaries[3].

To the best of our knowledge, none of the major NLP conference proceedings or journals currently provide lay summaries, or the equivalent. In implementing this mechanism for the first time in this field, we suspect some experimentation will be needed to set guidelines and best practices, and initially, we recommend not being overly prescriptivist as to the form of the lay summaries (in terms of length, format, content, etc.).

AI Safety Disclosure The goal of this second rider is to provide a common mechanism for applicable papers to highlight possible failure cases, even if just in broad terms–and even if in a relatively succinct format. Such error scenarios are not always obvious to downstream implementers, and the insight of the original researchers on the behavior of a system can, we surmise, often yield useful general guidelines for future work to consider. A description of failure cases can include an empirical analysis of inputs that generate incorrect or otherwise unreliable or uncertain outputs, but will often be of a more general, qualitative nature, highlighting potential biases in the output and future work needed to ensure reliable effectiveness in a real-world deployment.

Recognition of error cases can ground researchers in the current state of approaches and provide insights for future research. "It's in the errors that systems make that it's most evident that they have not cleared Turing's hurdle; they are not 'thinking' or 'intelligent' in the same sense in which people are" (Grosz, 2012). At the same time, analyzing errors in a systematic, representative fashion is non-trivial, and the next step of providing interpretable insights is perhaps harder still, and the subject of a burgeoning literature (Doshi-Velez and Kim, 2017). Simply asking researchers to highlight challenges in interpreting their models and problem cases in real-world deployments does not, of course, directly in itself yield innovations in error analysis or model interpretability. However, it does, we believe, encourage researchers to pay additional attention to these issues, and importantly, yields useful guides for downstream work.

Unlike lay summaries, the idea of an AI safety disclosure does not have an exact parallel in other fields nor existing mechanisms in the computer science publishing regime. It is in the spirit of existing guidelines for the treatment of human subjects in research, such as the Collaborative Institutional Training Initiative (CITI) Program, and the basic ethical principles of the Belmont Report

[1] http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-3806/homepage/ForAuthors.html#_Lay_Summary (accessed March 2018)

[2] http://blog.pnas.org/iforc.pdf (accessed March 2018)

[3] The list is available at https://elifesciences.org/inside-elife/5ebd9a3f/plain-language-summaries-journals-and-other-organizations-that-produce-plain-language-summaries (accessed March 2018)

(National Commission, 1979); however, importantly, it would also involve cases that would not typically be subject to review by an institutional review board (IRB). Existing IRB procedures are already well-suited for their target use-cases, and the AI safety disclosure is by no means intended to replace such mechanisms. On the contrary, we recommend that the AI safety disclosure be introduced as a voluntary endeavor with initially relatively informal guidelines, allowing the community to establish best-practices in a bottom-up fashion. In this sense, it is a much lighter-weight alternative to (and largely orthogonal to) the creation of ethics review boards for non-university organizations (Leidner and Plachouras, 2017) and is not intended to involve any particular additional legal obligations.

Perhaps more so than lay summaries, this second type of rider is likely going to need several iterations of experimentation before the community converges on standard guidelines. Given the heterogeneity of papers in NLP, it may well turn out that a single format is not suitable for all types of applicable papers in the field. In Section 4, we propose that ACL workshops can serve as useful testing grounds toward this end.

3 Example

We take a recent paper on the user-facing task of sentence correction (Schmaltz et al., 2017) and provide an example of a Lay Summary and an AI Safety Disclosure.

Lay Summary *This paper presents an approach for automatic grammar correction. The model for correction is based on models shown in previous work to be useful for the related task of automatic translation between languages, such as from Chinese to English. These types of models are referred to as a sequence-to-sequence models and are a type of neural network. The paper demonstrates ways of adapting these translation models for use in automatically correcting the grammar of English sentences. Effectiveness is improved over some previously proposed approaches, but the models are still noticeably worse than humans at the task.*

AI Safety Disclosure *Effectiveness at the demonstrated levels likely falls short of what is needed for a production system, but ensembles of models (including the intersection of language models) may increase effectiveness. However, since a non-zero proportion of the end users of such a system would likely be minors, it is worth mentioning some general principles to keep in mind when building such a system. In particular, a system built in the manner proposed here would not be particularly robust against biases already present in the aligned parallel data. Flipping of gendered pronouns may occur, and phrases offensive to at-risk populations could be generated. While not explored in the current work, an additional, final classifier may be helpful in filtering such changes.*

Learners might be sensitive to errors generated by such a system, learning to emulate the mistakes made in the output of the system. Without additional outside feedback and instruction, humans might learn to make the same false-positive and false-negative errors that the system makes. There is also a larger question of whether the existence of strong automatic correction systems will have the unintended effect of being detrimental to language learning, as students may become overreliant on such tools. This, too, needs to be investigated further.

4 Implementation

In order to minimize disruption of existing peer-review practices and establish best-practices, we recommend that the use of the two riders first be tested in a workshop setting. Additionally, we recommend that the riders not be included as part of papers during peer-review and remain voluntary.

We suspect that adoption of this proposal will be closely correlated with both the real and perceived amount of additional time required on the part of researchers. This can be partially alleviated by providing a series of examples using existing, published papers; however, as with other aspects, we want to emphasize that a goal is to not be overly prescriptivist and to allow the community to establish practices in a bottom-up, decentralized fashion.

Perhaps the most significant administrative effort will need to be placed in deciding how to make these riders accessible to the public. There are, for example, a variety of approaches in how existing science journals present lay summaries (Shailes, 2017), and we defer to conference and journal administrators on how best to present these riders.

5 Challenges

The proposal here is a modest departure from what already exists in other fields and is proposed as a voluntary endeavor. However, as with any policy proposal, there will be both anticipated and unanticipated downsides, and we briefly consider the possibilities here.

In terms of lay summaries, it is not a forgone conclusion that all researchers will be able to provide a summary that is understandable by a general audience. Of note, the current *PNAS* guidelines follow an earlier experiment with longer one-to two-page summaries, which "proved a burden for authors and editors. Some authors hit the mark precisely, but more frequently, the summary did not convey the salient features of the paper for a nonexpert" (Verma, 2012). Writing a summary for a general audience is non-trivial but learnable (Dubé and Lapane, 2014), and to the extent that computational tools can assist authors, the NLP community is in a unique position to develop such tools. While not a goal of this proposal, it is possible that a focus on such lay summaries could spark the development of tools that would be of use to authors in other areas of science, as well.

With the AI safety disclosure, we may find that in practice, the disclosures for some common tasks will be very similar across papers. It is possible that including this rider may become a mechanical exercise, with a small set of points reproduced across papers. It is possible that in such a scenario, the riders would be simply ignored in most cases by readers and authors, alike. One way to avoid this outcome would be to create an evolving challenge set of inputs/scenarios for common tasks on which previous approaches fail. The disclosures could then include results on these common sets, as well as announce additions to the challenge sets.

Researchers may be reluctant to acknowledge the potential downsides of their research. In some cases, a conflict of interest may prevent fully disclosing negative impacts. One approach to displaying the riders would be to do so with a forum that allows feedback from fellow researchers, perhaps in the style of the public reviews of openreview.net. However, invariably, there will be unevenness in the quality of the riders provided by authors (and/or in subsequent feedback), and the community will have to decide whether the benefits of having such riders outweigh such inconsistencies.

As noted above, these riders are not intended to carry any additional, particular legal weight (beyond that already present in the current research and publishing regime) in preventing a downstream application from implementing a system in contravention of concerns raised in a given "AI Safety Disclosure". However, we surmise that this type of bottom-up, public, decentralized approach can often be quite effective in influencing community norms.

6 Related Work

There is an emergent literature on AI safety and research ethics. Hovy and Spruit (2016) sparked recent research on the ethical significance of NLP research, with a focus on the impact of NLP on social justice. The contemporaneous work of Gebru et al. (2018) proposes a common mechanism for specifying potential biases within, and other characteristics of, datasets and trained models. The resulting "datasheet" is in the spirit of, and compatible with, our proposal, and in future work, we plan to explore combining these approaches. Grosz (2018) notes that "ethics must be taken into account from the start of system design", and the proposal here might be one small step in encouraging researchers to consider broader ethical implications as they develop their research.

There is a related, older literature addressing the limitations and potential unintended societal risks of complex, high-impact computational systems, more generally, of which the analysis of command and control systems is an illustrative example (Borning, 1987). A common theme of such work, as in the more recent work on biases in training data, is that data and technology reflect the social and political zeitgeist in which they are constructed. Technological solutions that ignore such coupling–even if well-intentioned–risk exacerbating existing tensions and creating new tensions.

There are a growing number of calls from scientists and journal editors for the need for lay summaries (Rodgers, 2017; Kuehne and Olden, 2015). Similarly, there is growing recognition for the need to both inform the general public about the state and possible future of AI, and to receive feedback from the public as stakeholders. Many of the realistic, near-term downsides of the current progress of AI, more generally, are likely to disproportionally impact those that are not AI researchers: commercial drivers, manufacturing workers, those in conflict zones, and those living under authoritarian governments, among others. Efforts to engage the public and/or broader cross-disciplinary collaborations include multi-disciplinary conferences, such as the recent AAAI/ACM conference on Artificial Intelligence, Ethics, and Society; public outreach efforts by organizations such as the Future of Life Institute[4]; and efforts to summarize progress in AI for a wider audience, such as the AI Index[5] (Shoham, 2017).

7 Conclusion

We recommend that future NLP papers presenting user-facing systems or methods include a short summary accessible to a general public and a brief overview of possible failure scenarios (even if speculative) of which future implementations and work should be aware. This proposal is a modest departure from what already exists in other scientific fields and involves a relatively lightweight change to existing publishing procedures in NLP. Experimentation of such an approach in an ACL workshop setting will be useful for gaining feedback from the research community and the public, and we recommend such an incremental, evaluative approach before applying it to full conferences and journals.

Acknowledgments

This paper benefited from discussions with colleagues at Harvard University and feedback from the anonymous reviewers. All errors are those of the author.

References

Alan Borning. 1987. Computer System Reliability and Nuclear War. *Commun. ACM*, 30(2):112–131.

Stanley F. Chen and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

F. Doshi-Velez and B. Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints*.

Catherine E. Dubé and Kate L. Lapane. 2014. Lay Abstracts and Summaries: Writing Advice for Scientists. *Journal of Cancer Education*, 29(3):577–579.

T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumeé, III, and K. Crawford. 2018. Datasheets for Datasets. *ArXiv e-prints*.

Barbara J. Grosz. 2012. What Question Would Turing Pose Today? *AI Magazine*, 33(4):73–81.

Barbara J. Grosz. 2018. Smart Enough to Talk With Us? Foundations and Challenges for Dialogue Capable AI Systems. *Computational Linguistics*, 44(1):1–15.

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.

Lauren M. Kuehne and Julian D. Olden. 2015. Opinion: Lay Summaries Needed to Enhance Science Communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

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.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

The National Commission. 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. *United States, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*.

Peter Rodgers. 2017. Plain-language Summaries of Research: Writing for different readers. *eLife*, 6:e25408.

Allen Schmaltz, Yoon Kim, Alexander Rush, and Stuart Shieber. 2017. Adapting Sequence Models for Sentence Correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2813, Copenhagen, Denmark. Association for Computational Linguistics.

Sarah Shailes. 2017. Plain-language summaries of research: Something for everyone. *eLife*, 6:e25411.

Yoav Shoham. 2017. Towards the AI Index. *AI Magazine*, 38(4):71–77.

[4] https://futureoflife.org/

[5] https://aiindex.org/

Inder M. Verma. 2012. PNAS Plus: Refining a Successful Experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34):13469–13469.