Linguistic Society of America
CENTENNIAL | 1924-2024

Incorporating ethical considerations directly in to the peer review process

Emily M. Bender
University of Washington

Overview

- Not ethics of reviewing, but reviewing for ethical considerations
- Societal impacts of language technology
- Recent history of field of compling grappling with ethical implications
- Case studies
- Implementation of ethical review in conference reviewing since 2020
- Reflections

Societal impacts of language technology, in brief

- Language models pick up bias from training data & amplify it
    - Biases come out in search engine results (Noble 2018), machine translation output (Caliskan & Lewis 2021), & more
- Systems trained for prestige varieties don’t work as well for non-standard/ stigmatized varieties (e.g. Wassink et al 2022, Sap et al 2019)
- … but better coverage of language varieties enables more surveillance (Bender & Grissom II, 2024)
- Data annotation processes often involve exploitative labor practices (Fort et al 2011)

Compling/NLP begins to grapple with this…

- Fort et al 2011: “Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine?”
- Éthique et Traitement Automatique des Langues, Journée d'étude de l'ATALA Paris, France, November 2014
    - Traitement Automatique des Langues 2016 Volume 57 Numéro 2
- Hovy & Spruit 2016: “The Social Impact of Natural Language Processing”
- Workshop on Ethics in Natural Language Processing at EACL (2017)
- => Regular track in ACL conferences

Case study #1: GERMEVAL 2020 Shared Task

- Originally: “Prediction of Intellectual Ability and Personality Traits from Text”
- Updated title: “Regression of artificially ranked cognitive and motivational style”
- Pushback on the corpora mailing list & twitter, but the task wasn’t rescinded
- Instead, a panel discussion was added to the conference
- More details:
    - https://bit.ly/GERMEVAL1
    - https://bit.ly/GERMEVAL2

Is this a problem for IRBs?

- People/communities affected often don't meet definition of “human subject”
- Not all compling research takes place a federally funded universities
- Even when it does people often just ... don't do IRB review

ACL takes action:
Adopting the ACM Code of Ethics

March 5, 2020

(PASSED)

ACL adopts the ACM Code of Ethics (https://www.acm.org/code-of-ethics) in the version adopted June 22nd, 2018, by the ACM Council. In its application to ACL, it is to be read in the contextually appropriate interpretation, e.g., "ACM member" is to be read as "ACL member". Sec 4.2 should be read as follows

"4.2 Treat violations of the Code as inconsistent with membership in the ACL. Each ACL member should encourage and support adherence by all members of the CL/NLP community regardless of ACL membership. ACL members who recognize a breach of the Code should consider reporting the violation to the ACL, which may result in remedial action."

ACM Code of Ethics

Ethics review at ACL conferences

- EMNLP 2020: Chairs Karën Fort, Dirk Hovy, plus 6 committee members
- NAACL 2021: Chairs Karën Fort, Emily M. Bender, plus 39 committee members
    - More lead time
    - Provided guidance to authors and reviewers
    - Ethics review for all flagged papers, as an educational strategy
    - Additional space allotted for “ethical considerations” sections

NAACL 2021 Ethics Review questions

- **For papers presenting new datasets:**
  - Does the paper describe how intellectual property (copyright, etc) was respected in the data collection process?
  - Does the paper describe how participants' privacy rights were respected in the data collection process?
  - Does the paper describe how crowd workers or other annotators were fairly compensated and how the compensation was determined to be fair?
  - Does the paper indicate that the data collection process was subjected to any necessary review by an appropriate review board?

NAACL 2021 Ethics Review Questions

- **For papers presenting new datasets AND papers presenting experiments on existing datasets:**
  - Does the paper describe the characteristics of the dataset in enough detail for a reader to understand which speaker populations the technology could be expected to work for?
  - Do the claims in the paper match the experimental results, in terms of how far the results can be expected to generalize?
  - Does the paper describe the steps taken to evaluate the quality of the dataset?

NAACL 2021 Ethics Review Questions

- **For papers concerning tasks beyond language-internal matters:**
- Does the paper describe how the technology would be deployed in actual use cases?
- Does the task carried out by the computer match how it would be deployed?
- Does the paper address possible harms when the technology is being used as intended and functioning correctly?
- Does the paper address possible harms when the technology is being used as intended but giving incorrect results?
- Does the paper address possible harms following from potential misuse of the technology?
- If the system learns from user input once deployed, does the paper describe checks and limitations to the learning?
- Are any of the possible harms you’ve identified likely to fall disproportionately on populations that already experience marginalization or are otherwise vulnerable?

NAACL 2021 Ethics Review Questions

- **For papers using identity characteristics (e.g. gender, race, ethnicity) as variables:**
  - Does the paper use self-identifications (rather than attributing identity characteristics to participants)?
  - Does the paper motivate the range of values used for identity characteristics in terms of how they relate to the research question?
  - Does the paper discuss the ethical implications of categorizing people, either in training datasets or in the deployment of the technology?

NAACL 2021 Ethics Review Process

- Primary reviewers flag papers for ethics review based on the ethics review questions

- The ethics committee recommends, for each flagged paper whether to accept as is, reject on ethical grounds (with explanation), conditional accept (with specification of what must be addressed).

- Camera-ready versions of papers designated as conditional accept are re-reviewed by the ethics committee to determine whether the concerns have been adequately addressed.

- The ethics committee is available to respond to questions from authors about the feedback they have received. This goes both for papers that were not accepted (for ethical reasons or otherwise), papers accepted as is, and papers conditionally accepted. In the latter case, we are happy to discuss during the preparation of the camera ready papers.

Current (ACL Rolling Review) Ethics Review Process

- Separate ethics reviewing step
- Ethics reviews must now ground issues identified in the ACM Code of Ethics
- Guidelines and examples: https://aclrollingreview.org/ethicsreviewertutorial

Case study #2: Prison Term Prediction

- Chen et al 2019: “Charge-Based Prison Term Prediction with Deep Gating Network”
  - Input: “accusation by the procuratorate” (p.6363) + corresponding set of charges (extracted by regex)
  - Output: the prison terms, in months, associated with the charges in the input
  - Data source: published records of the Supreme People’s Court of China
  - Intended use: providing an independent check in a phase of the proceedings where judgments are reviewed

Case study #2: Prison Term Prediction

- Leins et al 2020: “Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?”
  - Analysis with Chen et al 2019 as a case study
  - Data ethics concerns: What happens if a case is voided, but persists in the derived data set?
  - Use case concerns: “It is arguable that decisions regarding human freedom, and even potentially life and death, require greater consideration than that afforded by an algorithm, that is, that they should not be used at all.”

Case study #2: Prison Term Prediction

- Tsarapatsanis & Aletras 2021: “On the Ethical Limits of Natural Language Processing on Legal Text”
  - “academic freedom” should be considered as a value in any decisions, and balanced against e.g. privacy considerations of data subjects
  - the diversity of value systems represented within the global NLP community means that for any particular issue (though they use privacy as an example), the community should default to the most permissive position
  - the “threat of moralism” in NLP: “the belief that substantive ethical values, other than the disinterested pursuit of knowledge for its own sake, should be integral goals of research.” (p.3597)

Is this censorship?
What about academic freedom?

- Publication venues are in the business of judging academic quality
- Ethical practice and scientific validity are almost never at odds
- With academic freedom comes responsibility for academic integrity (see also Andy Perfors' blog post)
- More details:
  - https://bit.ly/ETHRV1
  - https://bit.ly/ETHRV2

Summary

- Pre-publication ethics review complements IRB review
- Publication venues are within their rights to judge the ethical suitability of candidate publications
- Ethics review processes should serve an educational function as well as any gate-keeping function

References

Bender, E. M. and Grissom II, A. (2024). Power shift: Toward inclusive natural language processing. In *Inclusion in Linguistics*. Oxford University Press.

Caliskan, A. and Lewis, M. (2021). Social biases in word embeddings and their relation to human cognition. In Dehghani, M. and Boyd, R., editors, *The Handbook of Language Analysis in Psychology*. Guilford Press.

Chen, H., Cai, D., Dai, W., Dai, Z., and Ding, Y. (2019). Charge-based prison term prediction with deep gating network. 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 6362–6367, Hong Kong, China. Association for Computational Linguistics.

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

Hovy, D. and Spruit, S. L. (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.

Leins, K., Lau, J. H., and Baldwin, T. (2020). Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Tsarapatsanis, D. and Aletras, N. (2021). On the ethical limits of natural language processing on legal text. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.