Understanding Ethics in NLP Authoring and Reviewing

Luciana Benotti and **Karën Fort** and **Min-Yen Kan** and **Yulia Tsvetkov**
acl-ethics-committee@inria.fr

Abstract

With NLP research now quickly being transferred into real-world applications, it is important to be aware of and think through the consequences of our scientific investigation. Such ethical considerations are important in both authoring and reviewing. This tutorial will equip participants with basic guidelines for thinking deeply about ethical issues and review common considerations that recur in NLP research. The methodology is interactive and participatory, including case studies and working in groups. Importantly, the participants will be co-building the tutorial outcomes and will be working to create further tutorial materials to share as public outcomes.

1 Motivation and structure

In late 2021, the Association for Computational Linguistics' executive committee appointed an Ethics Committee to investigate long-term ethical issues of the community's research and legislate any policy and workflow changes to the authoring, reviewing and other processes. The committee surveyed the constituency's opinions, wants and needs, finding that the majority of respondents felt that clear guidelines on acceptable practices regarding authoring and reviewing were needed. Specifically, in response to the question "What do you think are the most urgent tasks for the global *CL ethics committee?", 50% of respondents highlighted the need for more resources and discussion forums to raise awareness in the community about ethical issues in research and to clarify ethical review policies, 36% specifically mentioned the importance of creating dedicated training materials for authors and reviewers, and 26% encouraged more outreach initiatives to facilitate discussion about ethical research in the community.

This tutorial proposal thus follows from the mandate from the survey, such that more interactive opportunities exist to best communicate and train our membership on ethical guidelines and research practices.

The tutorial also draws on related, successful past tutorials on NLP reviewing and socially responsible NLP (≈100 participants) (Cohen et al., 2021; Tsvetkov et al., 2018), where some of the proposed tutorial instructors have been involved.

We propose a hybrid tutorial to best allow equitable access to the topic of this tutorial, especially to familiarize new community members and those who cannot afford access to attend physically. We plan to have dedicated presenters that can coordinate activities for the expected online participants. We may plan to use specific e-resources that can help facilitate virtual group discussions (e.g., Padlet, PollEverywhere, Google Docs, Slack).

We intend to make the tutorial presentation materials publicly available, in alignment with the stated goals of the tutorials. As an example, annotated presentation slides (with presenter notes) will be made available, such that tutorial participants can bring exercises of different lengths into classroom settings for research groups as well as undergraduate and graduate classes. We will organize a separate website via a Github repository[1] (to be owned by the ACL) to centralize our tutorial resources for long-term and public access.

However, due to the sensitive and formative nature of the small-group discussions, we will not record the small-group discussions so that participants can speak freely and off-the-record. The plenary, lecture-styled sessions (Sessions 1 and 7) may be recorded live, or pre-recorded offline.

This proposal tutorial aligns with the theme track "Reality Check"of ACL 2023. Most of the challenges addressed by the theme track, including out-of-domain generalization, adversarial attacks, spurious patterns (both linguistic and social), insensitivity to basic linguistic perturbations such as

[1] https://github.com/acl-org/ethics-tutorial, or similar (not yet published).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics
Tutorial Abstracts, pages 19–24

Segment Topic	Led by
1. Introduction and Foundations for Ethics	Presenters
2. Case Studies: Problematic Ethical Research — First reading	Participants
3. Structured Interaction / Dialogue	Presenters, Participants
4. Case studies — Second reading (Rotation)	Participants
5. Group Presentations	Group Leads
6. Summary and Common Issues	Presenters
7. Discussing and Troubleshooting Ethics and Further Resources	Presenters

Table 1: Tutorial Outline. Each segments' duration is ∼30 minutes, but 3 hours in total. Segments 2–6 will be conducted in small-group interaction.

negation, sensitivity to perturbations that should not matter (e.g., order and wording of prompts), are deeply related to ethical considerations of NLP research. In particular, proper discussion of risks (e.g., failure modes and vulnerabilities to adversarial attacks) and limitations (the scope of your claims, not overselling) is an integral to the theme and also for ethics authoring and reviewing. Finally, the theme track raises the question "what is an improvement in the real-world?", which is directly related to the social impact issues addressed by ethics reviewing.

2 Tutorial Content

Type: 1/2 day, Introductory
Expected Attendees: 100
Audience: Authors and reviewers, interested parties
Desired Location: Preferably ACL (Toronto, Canada)
Prerequisites: Introductory background in natural language processing and deep learning, including a basic familiarity of commonly-used approaches to text classification and generation, and standard NLP tasks. Fluent command of English.

Ethical consideration overarch our duties as researchers and scientists. As members of our community, and representatives of our works to both the general public and practitioners, we need to consider the ramifications of our work. The need for a better understanding of ethics is reflected in both authoring and reviewing, key functions of our community's peer review process.

Unintended and harmful ethical lapses and consequences can be largely avoided through continuing communication. Rather than assume that research is purely an intellectual pursuit, our tutorial invites participants to consider ethics as an integral component of the holistic framework of impactful research work. Table 1 presents our proposed tutorial's outline. Our aim is to provide hands-on experience with ethical issues through a small-group activity, both at the physical conference and in breakout rooms for online participants.

Ethics requires healthy debate and deep thought, and for these reasons, our structure incorporates a Socratic exercise, where participants spend a large part of the session discussing a concrete case of problematic research. A Community of Inquiry[2] approach will be taken such that participants engage in role-playing and discussing about ethical issues through reading 1–2 problematic hypothetical research abstracts from a curated set (§ 2.1). Using Socratic-style questioning, presenters guide the participants to engender discussion and realise ethical issues in the works.

Importantly, the participants will be co-building the tutorial outcomes and will be working to create further tutorial materials to share as public outcomes of the exercises. For many issues in ethics, the evolving discussion creates more value than the actual conclusions. This is why we propose such a dialectic approach.

To encapsulate the exercise, the presenters will first introduce the key ways that ethics impacts authoring and reviewing (Segment 1), summarise the group discussions' key points (Segment 6) and conclude with pointers to references and other training materials (Segment 7), including best practices for authoring ethical consideration sections (Benotti and Blackburn, 2022) and reviewing.

Due to the necessary interactivity of the session, we plan to limit the registrations for the tutorial to 100. This is to cater to having approximately a 25:1 ratio for presenters to participants. A larger volume than this jeopardizes the necessary interactive nature of the tutorial, which requires input from all participants.

2.1 Case studies

In the interactive portion of the tutorial, we will discuss research abstracts and will facilitate group discussions guided by critical questions about the proposed technology. Participants will be encour-

[2] https://en.wikipedia.org/wiki/Community_of_inquiry

aged to discuss the following questions:

- Ethics of the research question: Would answering this research question advance science without violating social contracts? What are potentials for misuse?
- Social impact of the proposed technology and its potential dual use: Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effects on people's lives?
- Privacy: Who owns the data? Understanding the differences between published versus publicized data, understanding the concept of user consent, and thinking about implicit assumptions of users on how their data will be used.
- Bias in data: What are possible artifacts in data, given population-specific distributions? How representative is this data to address the target task?
- Social bias and unfairness in models: Is there sufficient control for confounding variables and corner cases? Does the system optimize for the "right" objective? Could the system amplify data bias?
- Is the proposed evaluation sufficient? Is there a utility-based evaluation beyond accuracy; e.g., measurements of false positive and false negative rates as measurements of fairness? What is "the cost" of misclassification and fault (in)tolerance?

Our case studies will be hypothetical; i.e., we will not use abstracts from existing studies but will create abstracts that will allow us to highlight potential ethical issues covering multiple, diverse ethics-related topics, including human subjects research and institutional review board (IRB) approval, bias and fairness, privacy, misinformation, toxicity/content moderation, energy considerations/green AI. We will develop several representative case studies for participants to choose from; we show an example below that illustrates multiple problematic aspects within one study, which was adapted from an actual problematic recent study.

The following abstract introduces an unethical research question, a demographically biased data set, a data collection procedure that violates user privacy, a problematic evaluation procedure, and claims/potential applications that can lead to significant harms to individuals.

> **Abstract:** Faces contain more information about sexual orientation than can be perceived by the human brain. We used deep neural networks to extract features from over 35 thousand facial images. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 80% of cases, and in 70% of cases for women. Accuracy increased to 90% and 80%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone detected with 55% and 53% accuracy for gay males and gay females, respectively. Such findings advance our understanding of the origins of sexual orientation and the limits of human perception. Given that organizations are using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

2.2 Readings

We will cover a diversity of primary research on ethics, sourced beyond the presenters' own works, in the plenary sessions of the tutorial. Also, due to the abbreviated length of the 1/2-day format, our tutorial will cross reference sources from the list, rather than specifically require participants to do readings before the tutorial.

A full reading list of over 200 works has been cross-compiled by the full ACL Ethics Committee, sourced from university courses on NLP Ethics and related topics. The list available on Github[3]. The list can be updated by pull requests and is sortable by both topic and publication type. Topics and readings include the following among others: 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 (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.,

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

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 Presenters (listed in alphabetical order)

Luciana Benotti (*luciana.benotti@unc.edu.ar*, she/her) is an Associate Professor at the Universidad Nacional de Córdoba, in Argentina. Her research interests cover many aspects of situated and grounded language, including the study of misunderstandings, bias, stereotypes, and clarification requests. She is the elected chair of the NAACL executive board and is also serving as a member at large of the ACL Ethics committee.

Karën Fort (*karen.fort@sorbonne-universite.fr*, she/her) is an Associate Professor at Sorbonne Université and does her research at LORIA in Nancy, France. She has been working on ethics in NLP since 2014. She was co-chair of the first two ethics committees in the field (EMNLP 2020 and NAACL 2021) and is co-chair of the ACL ethics committee. She has been a member of the Sorbonne IRB between 2019 and 2022 and she teaches ethics at undergraduate and graduate level in Paris, Nancy, and the University of Malta.

Min-Yen Kan (*kanmy@comp.nus.edu.sg*, he/him): Associate Professor at the National University of Singapore and a co-chair of the ACL Ethics Committee. He has taught over 5,000 graduate and undergraduate students on his research interests in digital libraries, information retrieval and natural language processing.

Yulia Tsvetkov (*yuliats@cs.washington.edu*, she/her) is an Assistant Professor at the Paul G. Allen School of Computer Science and Engineering at the University of Washington, USA. Her research focuses on computational ethics, multilingual NLP, and machine learning for NLP. She developed a course on Computational Ethics in NLP and is teaching it at both undergraduate and graduate levels since 2017, and she is a co-chair of the ACL Ethics Committee.

4 Diversity considerations

The instructors of this tutorial are affiliated in different geographic regions. Luciana Benotti is in Latin America, Kären Fort in Europe, Min-Yen Kan in Asia and Yulia Tsvetkov in North America. Three of them identify with the female gender and one with the male gender. All of them are part of the ACL Ethics committee. We will promote this tutorial to all the ACL members but in particular to affinity groups such as Masakane, LatinX, North Africans, disabled in AI, indigenous in AI, Khipu and similar groups with the help of EquiCL. EquiCL is the only Big Interest Group in the ACL, its scope is equity and diversity and its current officers are Marine Carpuat (chair), Aline Villavicencio (secretary), Zeerak Waseem (communication with workshops and affinity groups). We think it is crucial to reach a diverse audience for this tutorial.

5 Ethical considerations

We are well aware that we do not compose a perfectly diverse committee and commit to pay close attention to ensure all participants' points of views are faithfully acknowledged.

We decided to use synthetic case studies in the form of abstracts, rather than real and complete articles, in order to preserve the anonymity of the authors, to refrain from personal criticism, and to allow the participants to focus more on the discussion than on the reading. We will create a variety of abstracts, with different forms, exemplifying different ethical issues, however, they will not cover all the possible ethical issues in the domain. Finally, the synthetic case studies will be clearly identified as such.

References

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.

Luciana Benotti and Patrick Blackburn. 2022. Ethics consideration sections in natural language processing papers. To appear in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, December 2022, Association for Computational Linguistics.

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.

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.

Kevin Cohen, Karën Fort, Margot Mieskes, Aurélie Névéol, and Anna Rogers. 2021. Reviewing natural language processing research. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 14–16, online. Association for Computational Linguistics.

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.

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.

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.

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.

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.

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.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2018. Socially responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, Louisiana. Association for Computational Linguistics.

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

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.