Navigating Ethical Challenges in NLP: Hands-on strategies for students and researchers

Luciana Benotti and **Fanny Ducel** and **Karën Fort** and **Guido Ivetta** and **Zhijing Jin**

Min-Yen Kan and **Seunghun Lee** and **Margot Mieskes** and **Minzhi Li** and **Adriana Pagano**

acl-ethics-chairs@inria.fr

Abstract

With NLP research being rapidly productionized into real-world applications, it is important to be aware of and think through the consequences of our research. Such ethical considerations are important in both authoring and reviewing (e.g. privacy, consent, fairness, etc).

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. Participants will gain practical experience on when to flag a paper for ethics review and how to write an ethical consideration section, which will be shared with the broader community. Importantly, the participants will be co-creating the tutorial outcomes and extending 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, finding that many 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% highlighted the need for more resources and forums to raise community awareness on ethical issues in research and to clarify ethical review policies; 36% mentioned the creation of dedicated training materials for authors and reviewers, and 26% mentioned outreach to facilitate discussion about ethical research.

This tutorial proposal follows from the mandate from the survey, so more interactive opportunities are offered to communicate and train our membership on ethical guidelines and research practices. Our proposal is also endorsed by the current president of the ACL, Emily M. Bender, who described ethics as a cornerstone of the synchronous development of CL/NLP: "*we need to place ethics in the curriculum for all NLP students — not as an elective, but as a core part of their education*".

The tutorial also draws on successful past tutorials on NLP reviewing and socially responsible NLP (≈100 participants) (Cohen et al., 2021; Tsvetkov et al., 2018; Benotti et al., 2023), where a part of our proposed tutorial team 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., PollEverywhere, Google Docs, Slack).

We will extend our publicly-available tutorial presentation materials[1]. As an example, annotated presentation slides (with presenter notes) will be modified and 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. 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.

Our tutorial aligns with the theme of **Introductory to fields related to CL/NLP**, and crucially, fulfills the mandate of the constituency. It helps CL/NLP researchers understand ethical concerns and its theoretical and practical implications. It

[1] https://ethics.aclweb.org/tutorials/

Segment Topic	Led by
1. Introduction and Foundations for Ethics	Presenters
2. Case Studies: Research that requires ethics review — 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.

also contextualizes particular aspects of the "limitations and ethical consideration sections" in articles and statistics of ethical issues brought about in reviews, to serve as illustrations of broader scientific considerations.

2 Tutorial Content

Type: 1/2 day, Introductory
Expected Attendees: 50
Audience: Authors and reviewers, interested parties
Desired Location: Preferably ACL (Vienna, Austria), but we would be happy to run the tutorial at multiple locations.
Prerequisites: Introductory background in natural language processing and deep learning, including a basic familiarity of commonly-used approaches to text mining and generation, and standard NLP tasks. Fluent command of English.

Ethical considerations 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 research which requires deep ethics review. A Community of Inquiry[2] approach will be taken such that participants engage in role-playing and discussing ethical issues through reading 1–2 problematic hypothetical research abstracts from a curated set (§ 2.2). Using Socratic-style questioning, presenters guide the participants to engender discussion and realise ethical issues in the works.

Participants will gain practical experience on when to flag a paper for ethics review and how to write an ethical consideration section that will be shared with the broader community. 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 *for flagging papers for ethics* 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 Flagging Guidelines

Flagging papers for ethics review is a critical first step in the review process, handled by the general program committee reviewers during their assessments. Justification is required for any paper flagged, to assist the ethics reviewers in further

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

evaluation. While essential for research integrity, ethics reviews must be applied judiciously to avoid overburdening the process and potentially stalling valuable research.

Participants who are potential scientific reviewers will receive training on deciding whether to flag a submission for further ethics reviewing from the ethics reviewing team.

Based on previous ARR cycles, we have identified patterns leading to unnecessary flags. Highlighting these issues will help reviewers to more accurately determine when a paper truly needs a deeper ethics review. We will discuss **points to avoid** such as *flagging without justification*, *flagging for missing section(s)*, and *flagging for reasons that could be addressed with feedback of the technical review*. We will also address some common misconceptions about the flagging process, including *"all data-centric papers require ethics review"* and *"the use of human annotators always requires ethics review"*.

2.2 Case studies

Our tutorial also covers the core ethics reviewing process itself. 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 encouraged 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.

2.3 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[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., 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 Diversity considerations

The instructors of this tutorial are affiliated in different geographic regions. Luciana Benotti and Guido Ivetta are in the Americas, Karën Fort and Fanny Ducel in Europe, Min-Yen Kan and Minzhi Li in Asia. They also represent different career stages: PhD candidates, associate, and full professors. Four identify with the female gender and two with the male gender. Luciana Benotti, Karën Fort and Min-Yen Kan are all co-chairs 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 in order to foster rich discussions.

We will also continue to promote participation in this tutorial and its topics through Birds of a Feather (BoF) sessions at *CL conferences. We will encourage conference ethics chairs and reviewers to do the same. Luciana Benotti and Guido Ivetta (two of the authors of this tutorial) together with Jocelyn Dunstan (Latin American ACL member working on privacy) organized one such BoF during NAACL 2024 in June that was attended by approximately 40 ACL members and Malihe Alikhani, ARR ethics chair, will host another one at EMNLP 2024 in December.

4 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.

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

5 Presenters (listed in alphabetical order)

Luciana Benotti (*luciana.benotti@unc.edu.ar*, she/her) is an Associate Professor at the Universidad Nacional de Córdoba, Argentina. Her research interests include situated and grounded language, especially the study of misunderstandings, bias, stereotypes, and clarification requests. She was the past chair of the NAACL executive board and is a co-chair of the ACL ethics committee.

Fanny Ducel (*fanny.ducel@universite-paris-saclay.fr*, she/her) is a PhD candidate at the Université Paris-Saclay, France. She works on stereotypical biases in LLMs. She also teaches ethics to international NLP graduates.

Karën Fort (*karen.fort@loria.fr*, she/her) is a Professor at the Université de Lorraine, France. She has been working on ethics and teaching ethics in NLP since 2014. She was co-chair of the first two ethics committees in the field, in 2020 and 2021 and is co-chair of the ACL ethics committee.

Guido Ivetta (*guidoivetta@mi.unc.edu.ar*, he/him) is a PhD candidate at the Universidad Nacional de Córdoba, Argentina. His work focuses on language model calibration and biases. He teaches AI ethics to K–12 teachers.

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 was a previous member of the ACL executive committee, and the inaugural Information Officer, and a previous ACL Anthology Director.

Minzhi Li (*li.minzhi@u.nus.edu*, she/her) is a PhD candidate at the National University of Singapore. Her main research interest is socially aware NLP.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4509–4516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Luciana Benotti, Karën Fort, Min-Yen Kan, and Yulia Tsvetkov. 2023. Understanding ethics in NLP authoring and reviewing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–24, Dubrovnik, Croatia. 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.

6 Appendix

Here are examples of other abstracts our team is developing and may field during the actual tutorial.

Abstract: Do people feel better on when on vacation? We study the social media sentiment of families when on vacation and during normal periods. Using a snowball sampling method, we solicited participants active on various social media platforms ($N = 1,337$) to voluntarily disclose when and where they went on vacation. We then crawled our participants social media posts to determine whether specific time periods and destinations for vacations were correlated with higher levels of enjoyment. Initial analysis showed that such features do yield significant prediction performance improvement: for example, vacations taken close-by tended to have markedly higher satisfaction. Our model, VACAY-OK, accounts for such factors in novel gating network architecture, improving joint vacation period and sentiment detection by over 10% F_1. Our analysis validates patterns of interest: vacations with immediate family exhibit a bimodal distribution, and that ones without in-laws have a significantly higher satisfaction and detection rate. For robustness in identification, we collect and study the pattern and text of social media posts by parents, children, extended family members and friends of vacation-goers. To ensure both the accessibility and the reproducibility of our experiments, we have gathered the raw social media data, culled with permission from the original posters, and released it as a zip archive, available without moderation, in the footnote.

The above abstract illustrates challenges with minors giving consent and with individuals releasing confidential data unknowingly (details shared by friends).

Abstract: In this paper, we present the largest existing language model as of today (989 trillion parameters, dataset of 968 Pb), `BigBlue`. We trained it using not only freely available Web content, but also Web archives and publicly-accessible EPub versions of all the books of the commericial Amazonia bookshop web storefront. The best results were attained using reinforcement learning with human feedback (RLHF) coming from crowdworkers. The model improves the state-of-the-art in Natural Language Processing (NLP) in most tasks, including sentiment analysis (+0.02 F-measure), dependency syntax (+0.015 UAS), named-entity recognition (+0.05 F-measure), etc. `BigBlue` is available through an API on our company's website.

The above abstract illustrates some of the many issues around LLM, including carbon footprint.

Abstract: In this paper, we present the first language model developed for Fridonian, a regional language spoken in the South-East of Austrafrancia by more than 500,000 people. We detail the collection of the corpus, the creation of the model and its evaluation on basic NLP tasks. The corpus creation itself is an achievement, as it is well-known and documented (Birdoff, 2020) that Fridonian speakers are reluctant to allow the recording of their language, for fear it will be used against them by the country's Sidonian majority. However, we managed to record 100 hours of speech in different settings: family discussions, local authority meetings, ceremonies and traditional story telling contests. We used this dataset to train a language model which we evaluated on sentiment analysis and named-entity recognition tasks, with impressive results (resp. 0.56 and 0.75 in F-measure). The created language model (`FridoBERT`) and the dataset (`FridoSPEECH`) are freely available on GitHub.

The above abstract illustrates the lack of consideration of the speakers and their needs and context.

Abstract: Is sentiment analysis dependent on culture or only on language? In this paper we describe the collection and annotation of a billion-word dataset over 7 different Spanish dialects spoken in Latin America, annotated with fine-grained sentiment analysis cues. To the best of our knowledge, this is the largest Latin American dataset currently available. We used data from public groups in Telegram where locals talk about products and services. We evaluate the state-of-the-art sentiment analysis models on it, showing that their performance is significantly lower than the state-of-the-art results on well-known benchmark for Spanish. The dataset annotation took three months. We selected crowdworkers after a careful exam to ensure they were fluent in the target dialect. We payed them 4 US dollars per hour, a fair wage considering it was two times the minimum wage in these countries. Our sentiment classifier fine-tuned per dialect outperforms previous models on Spanish by large margins. Our model has been used for 12 months by hundreds of companies in Latin America. To serve the model, we implemented a costing model to provide this service at a uniform price to all regions of the world. We conducted three usability studies in collaboration with three companies, which demonstrated that the integration of our sentiment analysis detector markedly increased users' engagement with their marketing advertisements.

The abstract presents several potential ethical challenges in data collection, model performance, and commercialization. First, the dataset was collected from public Telegram groups, raising concerns about privacy and informed consent, as users may not have been fully aware their data was being harvested for sentiment analysis. Additionally, sentiment analysis models can be culturally dependent, and there is an ethical risk that the fine-grained cues might not accurately capture the nuances of different dialects, leading to biases or misrepresentation. Economic fairness is another issue, as the crowdworkers were paid 4 USD per hour, which, while considered twice the minimum wage in some regions, may still be viewed as exploitative given the global profits likely generated from their labor. Furthermore, the model's use by companies for marketing could raise concerns about manipulation of consumer behavior, especially if the model tailors advertising based on dialect and exploits specific cultural tendencies. Lastly, while the uniform global pricing model for the sentiment analysis service may appear equitable, it could disadvantage regions with lower purchasing power, perpetuating technological and economic disparities. Addressing these ethical challenges requires transparency in data use, fair compensation for labor, and minimizing biases in AI models across cultural contexts.

Abstract: With the rise of large language models (LLMs), they have become useful in critical settings such as healthcare support, helping reduce administrative burden and improve predictive analytics. However, there is an emerging concern that LLMs encode gender and racial biases, which propagate to healthcare decisions. In this work, we propose a simple and effective approach to controlling for biases in LLMs through data anonymization, specifically focusing on patient names that often reveal gender and race. We introduce a novel and highly granular dataset of over 9.9 million patients' electronic health records annotated with patient demographics, including gender and race. We first fine-tune state-of-the-art LLMs on the raw healthcare data, and establish the presence of harmful biases in standard NLP benchmarks such as coreference resolution (10% lower F1 score for women compared to men) and named entity recognition (7% lower for White compared to Black patients). We also show that naively fine-tuned language models can be used to predict gender and race of the patients in held-out health records with high accuracy. Then, we show that a simple anonymization of the health data, by replacing patient names with generic placeholders, reduces the prediction gap in finetuned LLMs by up to 4% for gender, and up to 2.5% absolute for race. Our findings address important questions for fairness in NLP and algorithmic decision-making. Our code and data are publicly available to facilitate reproducibility.

The above abstract illustrates privacy concerns with using healthcare data, releasing data publicly, the problem with assumptions about privacy – that obfuscating names does not anonymize data since many other features correlate with demographics, annotation issued with binary gender and race, US-centric assumptions about conceptualization of race, evaluation of fairness on intrinsic benchmarks that might not reflect the true use case scenarios. Additional potential discussion is on what constitues "high accuracy" when the harm from misclasification is high (the weight of wrong predictions should be high).