The Future of NLP Needs Ethical Considerations: An Interactive Tutorial for Researchers and Students

Xi Ai and **Vamshi Bonagiri** and **Monojit Choudhury** and **Abdoul J. Djiberou Mahamadou**

Esther Gan and **Zhijing Jin** and **Min-Yen Kan** and **Seunghun J. Lee**

Yisong Miao and **Margot Mieskes** and **Aurélie Névéol** and **Yajing Yang**

acl-ethics-committee@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 work. Such ethical considerations are important in both senior and junior roles of researchers and students (e.g. privacy, consent, fairness, among others).

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 discussion and group work related to case studies that involve bias and broader issues (e.g., environmental impact). Participants will gain practical experience on when to flag a paper for ethics review and how to write an ethical consideration section to be shared with the broader community. Most 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 ACL Ethics Committee (hereafter, AEC) to oversee long-term policy on ethical issues, and to charge it with training its constituent membership on ethical issues. The AEC carried out a 2023 survey of the constituency, finding that many respondents felt that clear guidelines on acceptable practices regarding authoring and reviewing were needed (Benotti et al., 2023a): half 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, and 26% mentioned outreach to discuss ethical research.

This tutorial proposal follows from the 2023 survey mandates, to provide interactive opportunities to communicate and train our membership on ethical guidelines and research practices. **Our proposal is also endorsed by the previous and current ACL presidents.** Both have 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*".

This proposal also draws on successful 2023 and 2025 past tutorials on NLP reviewing and socially responsible NLP (Cohen et al., 2021; Tsvetkov et al., 2018; Benotti et al., 2023b, 2025) that have attracted 50–80 attendees. Past members of our team have been involved in these iterations.

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 and plan to use e-resources that facilitate virtual group discussion (e.g., Google Docs).

Our tutorial aligns with the EMNLP 2026 theme of **New Missions for NLP Research**, in calling for a re-think of the integral incorporation of ethics in NLP evaluation. Our tutorials will also help to fulfill the strong mandate given by our constituency.

2 Tutorial Content

Type: 1/2 day, Introductory
Expected Attendees: 50–100
Audience: Researchers, students, and other interested parties
Desired Location: Preferably in **both** locations.
Prerequisites: Introductory background in natural language processing and deep learning. Familiarity with text mining and generation and standard NLP tasks. Fluent command of English.

Note about location: We have tutorial organizers ready to be be physically and virtually present for *both* AACL-IJCNLP (Hangqin, China) and

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 11–19
November 6-10, 2026 ©2026 Association for Computational Linguistics

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, a total of 3 hours (3 1/2 w/ break). Segments 2–6 will be conducted in small-group interaction.

EMNLP (Budapest, Hungary). We strongly hope that the tutorial chairs can consider allowing this tutorial to be delivered in both locations as:

· This tutorial has not been fielded at EMNLP, which reaches a different group of researchers;
· This tutorial has not been fielded in Asia, which is burgeoning seat of future NLP research.

Ethical considerations overarch our duties as researchers and scientists, and we need to consider the ramifications of our work. The need for a better understanding of ethics is reflected in both researcher and student roles.

Unintended and harmful ethical lapses and consequences can be largely avoided through continuous 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 Inquiryapproach 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 realize ethical issues in the works.

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). 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.

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 Topics (Segments 1 & 6)

Depending on the location(s) where the tutorial will be executed, we will run versions of the tutorial that best fit the expertise of the participating instructors. We detail some possible topics below.

Multilingual and Multicultural Bias. *Sample Presenters: Xi Ai, Vamshi Bonagiri, Abdoul J. Djiberou Mahamadou, Monojit Chowdhury, Zhijing Jin and Seunghun J. Lee.*

This section of the tutorial will take a multilingual and multicultural perspective on bias and equity in NLP, with particular attention to low-resource languages and under-represented communities. It will cover fairness-aware data practices and evaluation beyond capability scores, including robustness and reliability metrics and auditing frameworks, along with mitigation strategies that promote more equal model quality and reduced harms across languages and cultures.

Environmental and Societal Impact. *Sample Presenters: Margot Mieskes and Aurélie Névéol.* This section of the tutorial will take a broad approach at NLP system evaluation (especially with respect to Large Language Models — LLMs) by considering impact on users and on the environment.

Our community's experiments have a cost: The machines on which the experiments are run have to be built and require resources to run (e.g., electricity, water for cooling). Participants will learn

how to assess impact incurred by complex NLP systems at various stages of the software lifecycle, including training, distilling, inference.

Furthermore, the experiments have social implications because the language data being processed often raises questions related to user privacy, speaker representation, and bias related harm.

Ethical Considerations for Author and Reviewer roles.[1] *Presenters: Esther Gan, Min-Yen Kan, Yisong Miao, Margot Mieskes and Yajing Yang.*

Ethics review is a critical step in the review process, where flagging is handled by reviewers during their assessments. While essential for research integrity, flagging must be applied judiciously with appropriate justification, to avoid hindering the process and stalling valuable research.

Our tutorial will train potential scientific reviewers on how to make this decision: We will highlight identified patterns (based on previous ARR cycles) leading to unnecessary flagging, helping reviewers to more accurately determine when a paper truly needs a deeper ethics review. We will review common mistakes to avoid such as *flagging for missing section(s)*, and *flagging for reasons that could be addressed with feedback of the technical review*.

2.2 Case studies (Segments 2–5)

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 that will relate closely to the topics presented during Segment 1:

- Conflict of Interests (COIs): What constitutes a COI? How does one report it? What interactions could result members with different roles (reviewer and author)?

- 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, the concept of user consent, and 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 edge cases? Is the system optimized for misaligned objectives or 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?

> **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.

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

[1] We note that there is complementary tutorial proposal from the ARR Chairs that will touch on ethics flagging. We believe that our tutorial section will complement theirs.

problematic aspects within one study, which was adapted from an actual problematic recent study.

The abstract above 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.

2.3 Readings & Resources (Segment 7)

We will cover a diversity of primary research on ethics, sourced beyond the presenters' own works, in the final 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[2] of over 200 works has been cross-compiled by the AEC, sourced from university courses on NLP Ethics and related topics. Topics and readings relate directly to the topics to be presented, which include: **data usage** (Mieskes, 2017; Bender and Friedman, 2018; Gebru et al., 2021), **crowdsourcing** (Fort et al., 2011; Callison-Burch, 2014; Fort et al., 2014; Hara et al., 2018; Toxtli et al., 2021), **biases** (Blodgett et al., 2020; Fort et al., 2024; Seth et al., 2025), **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; Tan et al., 2021), **environmental impact** (Strubell et al., 2019; Zhou et al., 2020; Henderson et al., 2020; Bannour et al., 2021) and **human harms and values** (Hovy and Spruit, 2016; Leidner and Plachouras, 2017).

Our tutorial also benefits from its prior iterations. We will extend our publicly-available tutorial presentation materials[3] available on the ACL Ethics website. 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. While we plan to record the plenary, lecture-styled sessions (Sessions 1, 6 and 7), 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.

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

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

3 Diversity considerations

The instructors of this tutorial are affiliated in different geographic regions. Zhijing Jin in North America, Aurélie Névéol and Margot Mieskes in Europe; Vamshi Bonagiri, Monojit Choudhury and Abdoul Jalil Djiberou Mahamadou in the Middle East; Xi Ai, Esther Gan, Min-Yen Kan, Seunghun J. Lee, Yisong Miao and Yajing Yang in Asia. They also represent different career stages: PhD candidates, postdoctoral researchers, assistant, associate and full professors; Three are co-chairs of the AEC, two others are members of the AEC and one other is a student volunteer with the AEC. Five identify with the female gender and six with the male gender. We note that there are potentially more junior presenters that will be identified by the core investigators in due course.

We will promote this tutorial to all the ACL members but particularly to affinity groups such as Masakhane, LatinX, disabled in AI, indigenous in AI, and similar groups with the help of EquiCL Big Interest Group. It is crucial to reach a diverse audience for this tutorial to best foster rich discussions. We also aim to accommodate attendees needing special attention for registration or attendance.

4 Ethical considerations

We 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 participants to focus on the discussion than on the reading. We will create a variety of abstracts (c.f., Appendix), with different forms, exemplifying different ethical issues; however, they will not cover all the possible ethical issues in the domain.

5 Logistics

Our tutorial requires standard conference presentation equipment, including a projector, screen, and both presenter and audience microphones (where feasible), and reliable internet access (for remote participation).

6 Limitations

We are requesting to check whether there are restrictions on content that should be avoided to be compliant to the local venue (especially within China, the venue for AACL-IJCNLP).

7 Presenters (listed in alphabetical order)

Note: We have annotated each presenter with two-letter codes to indicate which venue ((E)MNLP or (A)ACL-IJCNLP) and modality of participation (**(P)hysical** or *(V)irtual*), where already known.

Xi Ai [E*V*, A**P**] (*baridxai@u.nus.edu*, he/him) is a research fellow at WING@NUS and NUSRI-CQ. His research focuses on multilingual and cross-lingual NLP and machine learning for NLP. He has co-organized Workshop on Computational Approaches to Linguistic Code-Switching (CALCS) at *CL conferences.

Vamshi Bonagiri [E**P**, A*V*] (*vamshi.bonagiri@mbzuai.ac.ae*, he/him) is a PhD student at MBZUAI. His research focuses on AI safety and the societal impacts of AI technologies. He has also taught and mentored various NLP and responsible and safe AI courses.

Monojit Choudhury [E**P**] (*monojit.choudhury@mbzuai.ac.ae*, he/him) is a Professor of NLP at MBZUAI, Chief Scientist of the Institute of Agriculture and AI, and a member of AEC. His work focuses on equitable creation and deployment of language technology.

Abdoul Jalil Djiberou Mahamadou [E*V*, A**P**] (*abdoul.mahamadou@mbzuai.ac.ae*, he/him) is a postdoc at MBZUAI working on personalization, cultural sensitivity, and inclusion in AI. He did a postdoc in AI ethics at the Stanford Centre for Biomedical Ethics, where he taught bioethics and research ethics. He researched algorithm bias, fairness, and participatory AI governance.

Esther Gan [E*V*, A**P**] (*esther.gan@u.nus.edu*, she/her) is a doctoral student with NUS and an active volunteer with the AEC. She has taught topics related to ethics as NUS teaching staff.

Min-Yen Kan [E*V*, A**P**] (*knmnyn@nus.edu.sg*, he/him) is an Associate Professor at NUS and a co-chair of the AEC. He was a member of the ACL executive committee, the inaugural Information Officer and a previous ACL Anthology Director.

Seunghun J. Lee [A**P**] (*seunghun@icu.ac.jp*, he/him) Senior Associate Professor of Linguistics at the International Christian University. He is also an Adjunct Professor of African Languages at the University of Venda in South Africa, and an Honorary Associate Professor in the Centre for Linguistic Science & Technology at the Indian Institute of Technology Guwahati, India.

Yisong Miao [E*V*, A**P**] (*yisong@comp.nus.edu.sg*, he/him) is a doctoral student at NUS. He has been an active member of the discourse area in ACL. He has been a teaching assistant at NUS for its NLP course, covering ethical topics.

Margot Mieskes [E**P**, A*V*] (*margot.mieskes@h-da.de*, she/her) Professor for Information Science at the University of Applied Sciences, Darmstadt, Germany. She works in both the CL and NLP areas, focusing on summarization, evaluation, data quality, ethics, experiment reproducibility, and biases in LLMs. She is also part of the ACL Rolling Review Editor in Chief team.

Aurélie Névéol [E**P**, A*V*] (*aurelie.neveol@lisn.fr*, she/her) is a Senior Researcher at CNRS and a co-chair of the AEC She has been a member of SIG BioMed since 2006. Her work addresses NLP applied to the biomedical domain guided by responsible science principles.

Yajing Yang [E*V*, A**P**] (*yajing.yang@u.nus.edu*, she/her) is a doctoral student at NUS where her research focuses on financial NLP and data-to-text generation. She was a teaching assistant for NUS' NLP course.

Zhijing Jin [E**P**, A*V*] (*zjin@cs.toronto.edu*, she/her) is an Assistant Professor at the University of Toronto. Her research focuses on LLMs, multi-agent system, causal inference in NLP with an emphasis on social good applications. She is a co-chair of the AEC and co-organized NLP for Positive Impact workshops at *CL conferences and the ACL Year-Round Mentorship.

References

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.

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, Mark Dredze, Karën Fort, Pascale Fung, Dirk Hovy, Jin-Dong Kim, Min-Yen Kan, Malvina Nissim, and Yulia Tsvetkov. 2023a. Yes, we care (more)! results of the 2021 ethics and natural language processing survey. Technical Report v230714, ACL Ethics Committee. Report on the 2021 Ethics and Natural Language Processing Survey, released 14 July 2023.

Luciana Benotti, Fanny Ducel, Karën Fort, Guido Ivetta, Zhijing Jin, Min-Yen Kan, Seunghun J. Lee, Minzhi Li, Margot Mieskes, and Adriana Pagano. 2025. Navigating ethical challenges in NLP: Hands-on strategies for students and researchers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 7–8, Vienna, Austria. Association for Computational Linguistics.

Luciana Benotti, Karën Fort, Min-Yen Kan, and Yulia Tsvetkov. 2023b. 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.

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.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanotto, and Aurélie Névéol. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

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.

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.

Agrima Seth, Monojit Choudhury, Sunayana Sitaram, Kentaro Toyama, Aditya Vashistha, and Kalika Bali. 2025. How deep is representational bias in llms? the cases of caste and religion. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2319–2330.

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.

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.

8 Appendix

We are unsure whether appendices are allowed, but we have included examples of other abstracts Here are examples of other abstracts our team is developing and may field during the actual tutorial. Please disregard if not allowed.

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).