Societal impacts of NLP:
How and when to integrate them into your research (and how to make time for that)

Emily M. Bender
University of Washington
@emilymbender

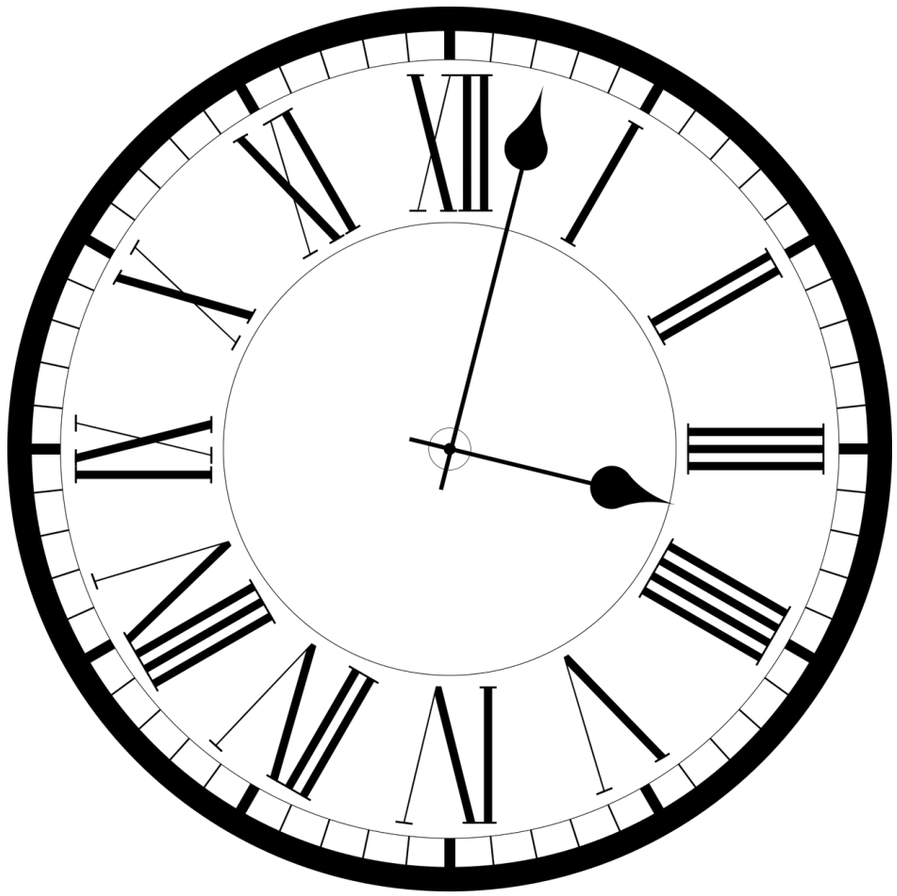
AACL-IJCNLP SRW
December 7, 2020

A talk about time…

- Locating current discussions about ethics and NLP (and ethics and AI more broadly) in historical perspective

- How to make time in your research practice for these considerations

- How to apply your valuable research time in the most beneficial ways



Recent ACL history

- ACL 2016 Hovy & Spruit "The Social Impact of Natural Language Processing"
- EACL 2017 First ACL Workshop on Ethics and Natural Language Processing
- NAACL 2018 Second ACL Workshop on Ethics and Natural Language Processing
- NAACL 2019 theme: The tension between data privacy and model bias in NLP
- ACL 2020 includes Ethics and NLP as a track

Recent ACL history

- March 2020, the ACL officially adopts the ACM code of ethics
- ACL/ACM code of ethics included in EMNLP 2020 call for papers

NEW: Ethics Policy

Authors are required to honour the ethical code set out in the ACM Code of Ethics. The consideration of the ethical impact of our research, use of data, and potential applications of our work has always been an important consideration, and as artificial intelligence is becoming more mainstream, these issues are increasingly pertinent. We ask that all authors read the code, and ensure that their work is conformant to this code. Where a paper may raise ethical issues, we ask that you include in the paper an explicit discussion of these issues, which will be taken into account in the review process. We reserve the right to reject papers on ethical grounds, where the authors are judged to have operated counter to the code of ethics, or have inadequately addressed legitimate ethical concerns with their work

Recent ACL history

- March 2020, the ACL officially adopts the ACM code of ethics
- ACL/ACM code of ethics included in EMNLP 2020 call for papers
- NAACL 2021 follows suit, but with enough lead time to provide guidance to authors on writing ethics statements
- EMNLP 2020 panel: Publishing in an Era of Responsible AI: How can NLP be Proactive? Considerations and Implications
- ACL-IJCNLP 2021 also includes the ethics policy
- (NAACL 2021 and ACL 2021 include tracks on “Ethics, Bias, Fairness” and “Ethics and NLP”)

Why do we need to do this?

- “The L in NLP means language, and language means people” (Schnoebelen 2017)

- We are building technology that affects people in the world, and we therefore have a responsibility towards those people
  - People whose data we collect
  - People who will use the technology
  - People who will be affected by others’ use of the technology

Towards a typology of risks of NLP: Guiding principles

- Value sensitive design (Friedman et al 2006, Friedman & Hendry 2019):
  - Identify stakeholders
  - Identify stakeholders' values
  - Design to support stakeholders' values
- Sociolinguistics (e.g. Labov 1966, Eckert & Rickford 2001):
  - Variation is the natural state of language
  - Status as 'standard' language is merely a question of power
  - Language varieties & features associated with marginalized groups tend to be stigmatized
  - Our social world is largely constructed through linguistic behavior

Stakeholder-centered typology

	Direct stakeholders	Indirect stakeholders
Tech user	User, by choice	Harm to community
	User, not by choice	Harm to individual
Tech dev	Annotator, crowdworker	Unwitting data contributor

(see also Hovy & Spruit 2016, Barocas et al 2017)

Direct stakeholders: Not by choice

- *My screening interview was conducted by a virtual agent*
- *I can only access my account information via a virtual agent*
- *Access to a emergency response system requires interaction with a virtual agent first*
  - … but it doesn't work or doesn't work well for my language variety
    - I scored poorly on the interview, even though the content of my answers was good
    - I can't access my account information or emergency response

Indirect stakeholders: Community harm

- *Systems are built using general webtext as a proxy for word meaning or world knowledge*
    - … but general web text reflects many types of bias (Bolukbasi et al 2016, Caliskan et al 2017, Gonen & Goldberg 2019; see also Blodgett et al 2020)
    - autocompletion of search queries repeats & reinforces harmful stereotypes (Noble 2018)

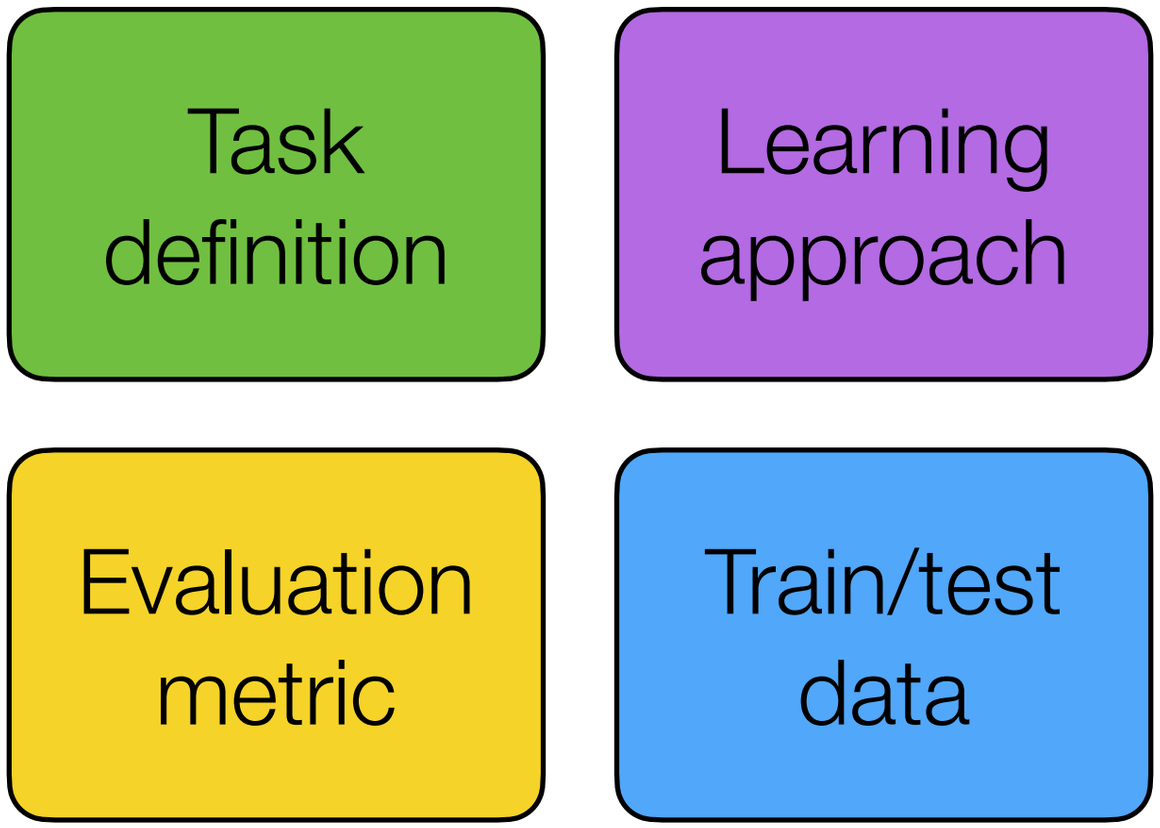
Explored in more detail at SSNLP 11 Dec 2020

What does this mean for NLP researchers & developers?

- We have a responsibility to broaden our lens:
    - our jobs aren't just about framing and solving technical problems
    - but also about understanding how the tech we build (or choose not to build) fits into society
- This requires a slower pace of "progress"
- Being systematic about documentation can help

Machine learning, in a nutshell

- “Each machine learning problem can be precisely defined as the problem of improving some measure of performance P when executing some task T, through some type of training experience E. […] Once the three components ⟨T,P,E⟩ have been specified fully, the learning problem is well defined” (Mitchell 2017, p.2)

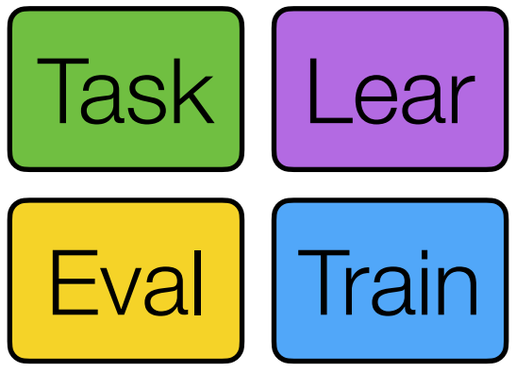


Machine learning, in context

Why do we care about this task?

How does dataset model the task?

-build something useful
-learn about: computers, people, modeling domain



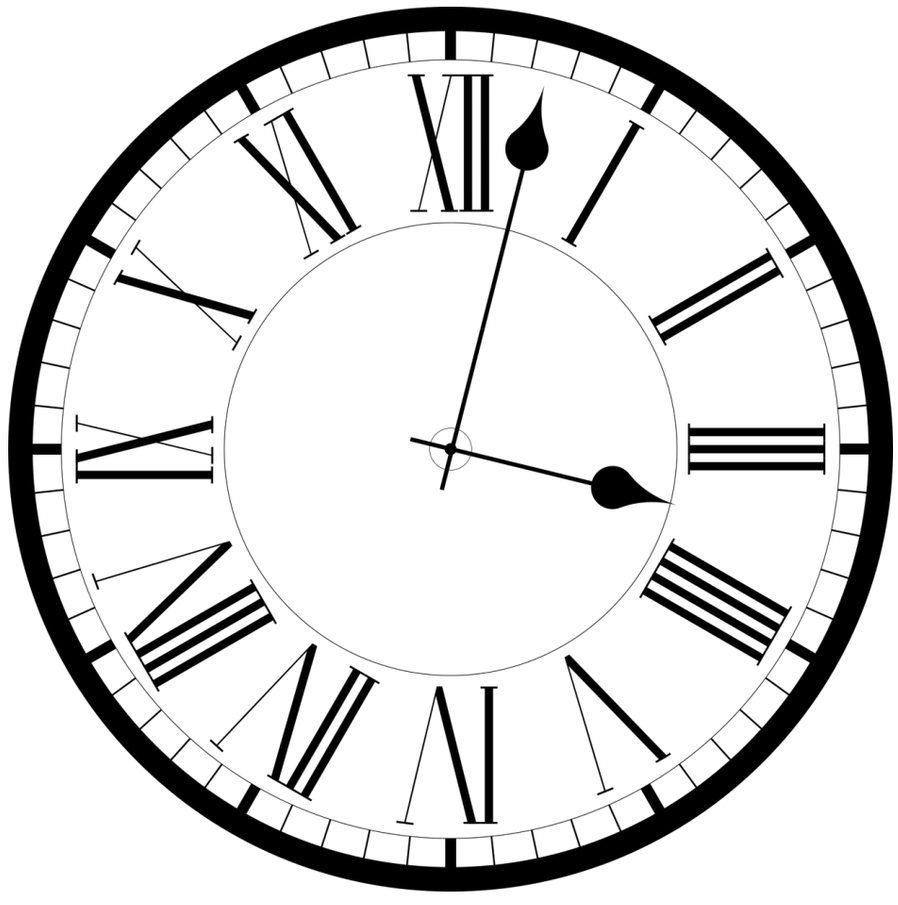
How does the task relate to the world?

What happens when we deploy this?

How do we collect the data?

Recent ACL history

- ACL adopts ACM code of ethics (March, 2020)
- EMNLP 2020, NACL 2021, ACL-IJCNLP 2021 include ethics policy in CFP
- Why do we suddenly have so many more responsibilities?



Beyond ACL

- Journée d'études ATALA "éthique et TAL" 2014, organized by Karën Fort and Benoît Sagot

- FATML -> FAT* -> FAccT since 2014: Fairness, Accountability and Transparency in Machine Learning

Brief history of scientific ethics (Metcalf 2014)

- 1940s: Nuremberg trials lead to World Medical Association's Geneva Declaration (1948) and Helsinki Declaration (1964)

- 1960s-1970s: Tuskegee, Willowbrook, Milgram, Stanford Prison and other abuses of research subjects lead to the Belmont Report and the establishment of ethics boards

Common principles of research ethics
(Metcalf 2014, p.2)

- respect for persons (autonomy, privacy, informed consent)
- balancing of risk to individuals with benefit to society
- careful selection of participants
- independent review of research proposals
- self-regulating communities of professionals
- funding dependent on adherence to ethical standards

How does this relate to NLP?

- Machine learning is often seen as “playing with data” and “solving abstract problems”
    - But: “The L in NLP means language, and language means people” (Schnoebelen 2017)
- Linguistics is not uniformly better: ethics board approval for child language work, psycholinguistic studies, and some documentary work
- So we need to do IRB review for NLP research?
    - Ethics boards tend to focus on preventing abuses of research subjects, not further downstream harms

Brief history of “Computer Ethics” (Bynum 2000)

- *Cybernetics: Or Control and Communication in the Animal and the Machine* (Wiener 1948)
- *The Human Use of Human Beings* (Wiener 1950)
- ACM Code of Ethics first adopted in 1973
- *Computer Power and Human Reason* (Weizenbaum 1976)
- 1983: Computer Professionals for Social Responsibility
- “What is Computer Ethics?” (Moor 1985)
- *Computer Ethics* (Johnson 1985)

Vision for NLP

- Engagement in discussions of ethics/societal impact that go beyond “this work is approved/unacceptable”
    - (Though some systems should not be built.)
- Working together to understand what potentials for harm there are, how to mitigate them, and how to educate the public about them
- Working together to develop best practices that help us do this work
- Move away from leaderboardism and towards work that is situated and interdisciplinary
    - (This requires a different, but healthier, time-scale.)

Ethical and scientific considerations are usually aligned!

Best practices

- Treat subjects fairly (informed consent, fair compensation)
- Abide by licenses and terms of use
- Data and model documentation (data statements, datasheets, model cards)
- Connecting model development work to specific problems in the world (with specific use cases and stakeholders)
    - Especially important in benchmark development
- Identifying stakeholders and, if possible, seeking their input
- Writing ethics statements which are proactive, rather than defensive
- Own your point of view … and learn from others'

Data Statements for NLP: Transparent documentation
(Bender & Friedman 2018)

- Foreground characteristics of our datasets (see also: AI Now Institute 2018, Gebru et al 2018, Mitchell et al 2019)

- Make it clear which populations & linguistic styles are and are not represented

- Support reasoning about what the possible effects of mismatches may be

- Recognize limitations of both training and test data:

  - Training data: effects on how systems can be appropriately deployed

  - Test data: effects on what we can measure & claim about system performance

Situating ML tasks in the world

- Make time to consider, early & often, the following questions:
    - What are the use cases of the technology being developed?
    - How does the specific ML task (inputs, outputs) relate to the intended use case?
    - What are the failure modes and who might be harmed?
    - What kinds of bias are likely to be included in the training data?
- Broaden our notion of ‘scaling up’: It’s not just about large numbers but also about diverse communities & experiences with the software

When writing papers

- Data collection: clearly document how the work adheres to license conditions, standards of informed consent, and norms of fair compensation
- Identify stakeholders, and describe possible risks
- Describe what might be done to mitigate those risks:
    - what information should be exposed about how systems work
    - how might regulation be informed
    - what does the general public need to know?
- Resist the pressure to “sell” your work as a perfect, world-saving system

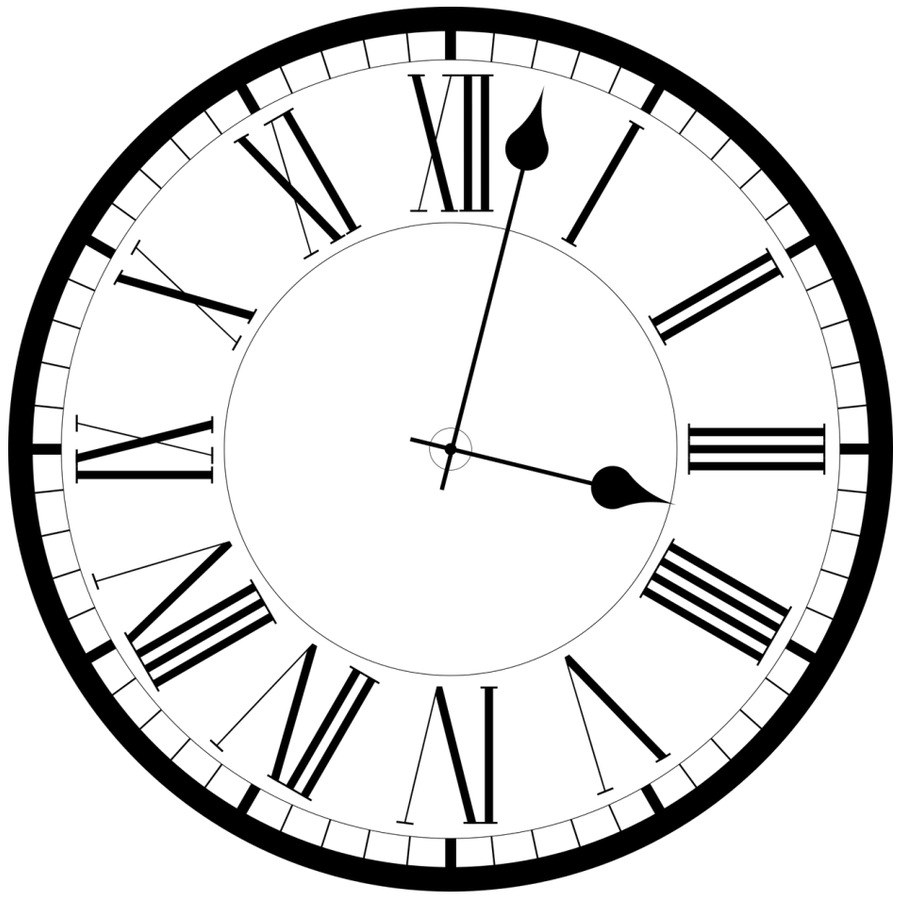
Own your point of view

- Norms of English-language scientific writing create a false and harmful veneer of “objectivity”
    - The “view from nowhere” doesn’t actually exist (Stitzlein 2004, Gebru 2020)
    - Adopting faux objectivity also makes it difficult to talk about harms
- Own your own point of view
    - This will help you describe possible pitfalls in your work
    - It will also help you challenge systems of power and oppression

How to make time for this

- Start here! Before committing time to model development for a task that may turn out problematic:
  - investigate the task
  - ensure fair and legal data collection
- Incorporate papers on ethics/societal impacts into reading groups
- Keep in mind: better not best, progress not perfection

Your time is valuable!!



How to do this as a student

- Keep a clear-eyed view of power relations
- Build networks of people to talk through concerns with
- Frame in terms of risks:
    - papers not getting accepted
    - institutions in legal trouble
- Progress, not perfection
- Ethical and scientific considerations are usually aligned!

Thank you!

Suggested reading

- www.acm.org/code-of-ethics
- Blodgett et al 2020 (ACL) "Language (Technology) is Power: A Critical Survey of "Bias" in NLP"
- Larson 2017 (EACL workshop) "Gender as a Variable in Natural-Language Processing: Ethical Considerations"
- Sweeney 2013 (CACM) "Discrimination in Online Ad Delivery"
- Noble 2018 *Algorithms of oppression: How search engines reinforce racism*
- Benjamin 2019 *Race after technology: Abolitionist tools for the New Jim Code*
- Agüera y Arcas, Mitchell and Todorov 2017 (medium.com) "Physiognomy's New Clothes"

+ Radical AI Podcast
www.radicalai.org

References

Agüera y Arcas, B., Mitchell, M., and Todorov, A. (2017). Physiognomys new clothes. Blog post on Medium.com, `https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a`.

AI Now Institute (2018). Algorithmic impact assessments: Toward accountable automation in public agencies. Medium.com.

Alfano, M., Hovy, D., Mitchell, M., and Strube, M., editors (2018). *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Barocas, S., Crawford, K., Shapiro, A., and Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In *SIGCIS Conference*.

Bender, E. M. and Friedman, B. (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.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (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.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Bynum, T. W. (2000). A very short history of computer ethics. *APA Newsletters on Philosophy and Computers*, 99(2):2.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Eckert, P. and Rickford, J. R., editors (2001). *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge.

Friedman, B. and Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Friedman, B., Kahn, Jr., P. H., and Borning, A. (2006). Value sensitive design and information systems. In Zhang, P. and Galletta, D., editors, *Human–Computer Interaction in Management Information Systems: Foundations*, pages 348–372. M. E. Sharpe, Armonk NY.

Gebru, T. (2020). Race and gender. In Dubber, M. D., Pasquale, F., and Das, S., editors, *The Oxford Handbook of Ethics of AI*. Oxford University Press, Oxford.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for datasets. arXiv:1803.09010v1.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Hovy, D., Spruit, S., Mitchell, M., Bender, E. M., Strube, M., and Wallach, H., editors (2017). *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Johnson, D. G. (1985). *Computer Ethics*. Prentice-Hall.

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Metcalf, J. (2014). Ethics codes: History, context, and challenges. Council for Big Data, Ethics, and Society. Accessed December 5, 2020. https://bdes.datasociety.net/council-output/ethics-codes-history-context-and-challenges/.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. ACM.

Mitchell, T. (2017). Machine learning, ch 14: Key ideas in machine learning. http://www.cs.cmu.edu/~tom/mlbook/keyIdeas.pdf.

Moor, J. H. (1985). What is computer ethics? In Bynum, T. W., editor, *Computers and Ethics*, pages 266–275. Basil Blackwell.

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

Schnoebelen, T. (2017). The carrots and sticks of ethical NLP. Blog post, https://medium.com/@TSchnoebelen/ethics-and-nlp-some-further-thoughts-53bd7cc3ff69, accessed 19 March 2019.

Stitzlein, S. M. (2004). Replacing the 'view from nowhere': A pragmatist-feminist science classroom. *Electronic Journal of Science Education*, 9(2).

Sweeney, L. (May 1, 2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. Freeman, San Francisco CA.