Societal Impacts of Language Technology: How to Work with Known Best Practices to Avoid Harm

Emily M. Bender
University of Washington
@emilymbender / @emilymbender@dair-community.social

PhD Masterclass
April 4, 2024
zhaw School of Applied Lingiustics

Goals

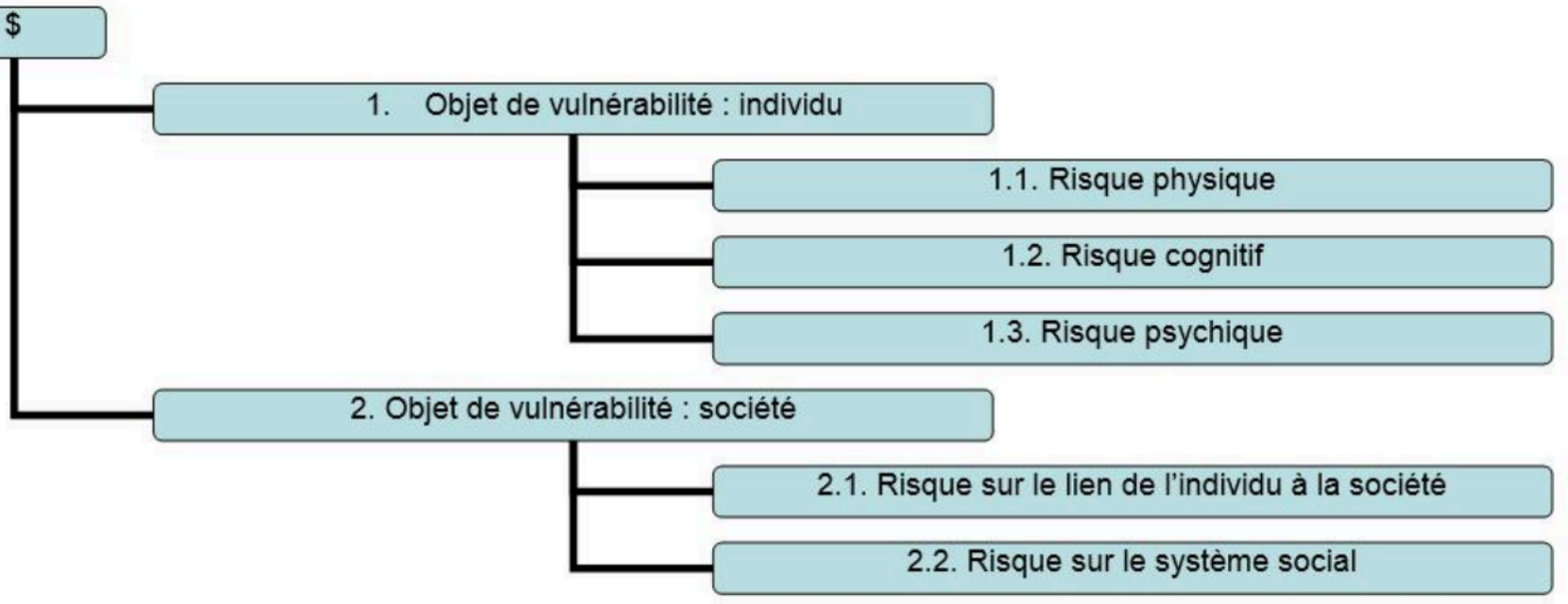
- Present a typology of the risks of adverse impacts of language technology
    - Grounded in value sensitive design and sociolinguistics
- Present a range of strategies for mitigating harm / minimizing risk
    - With a view towards “progress, not perfection”

Typology

- A systematic classification of phenomena, along one or more dimensions
- Helps to explore the space of possibilities
- Helps to understand relationships across categories

Typology of risks (Lefeuvre-Halftermeyer et al 2016)

Figure 2. *Deux premiers niveaux de la typologie : vulnérabilité et classe d'impact*



Hovy & Spruitt 2016
“The Social Impact of Natural Language Processing”

- Survey of some types of issues

- Importantly raised awareness of the discussion within English-language NLP circles

- Introduced concepts of:

    - Exclusion, Overgeneralization, Bias confirmation, Topic Overexposure, Dual use

    - Illustrated with NLP-specific examples of negative impacts

- Not exhaustive, not a typology

Taxonomy for algorithmic harm reduction (Shelby et al 2023)

Figure 1: Sociotechnical harms taxonomy overview.

- Stereotyping social groups
- Demeaning social grups
- Erasing social groups
- Alienating social groups
- Denying people the opportunity to self-identify
- Reifying essentialist social categories

- Opportunity loss
- Economic loss

- Alienation
- Increased labor
- Service/benefit loss

- Loss of agency
- Tech-facilitated violence
- Diminished health and well-being
- Privacy violations

- Information harms
- Cultural harms
- Civic and political harms
- Socio-economic harms
- Environmental harms

Another taxonomy of harms [Not mutually exclusive] (from Barocas et al, 2017)

- allocational harms: ML systems unfairly allocating finite resources
- representational harms: ML systems contribute to subordination of certain groups
  - quality of service (e.g. ASR working better for some groups than others; Koenecke et al 2021, Wassink et al 2023)
  - stereotyping (e.g. online ads suggesting that people with Black-sounding names had been arrested; Sweeney, 2013)
  - denigration (e.g. Tay, where the ML system actively participated in hate speech; Price, 2016)
  - under-representation (e.g. image search for “CEO” returning more images of white men than is reflected in the real world; Kay et al, 2015)

Guiding principles: Value sensitive design

- Value sensitive design (Friedman et al 2006, Friedman & Hendry 2019):
    - Identify stakeholders
    - Identify stakeholders' values
    - Design to support stakeholders' values

Guiding principles: Sociolinguistics
(e.g. Labov 1966, Eckert & Rickford 2001)

- Variation is the natural state of language
    - Variation in pronunciation, word choice, grammatical structures
- Status as ‘standard’ language is a question of power, not anything inherent to the language variety itself
    - Language varieties & features associated with marginalized groups tend to be stigmatized
- Meaning, including social meaning, is negotiated in language use
- Our social world is largely constructed through linguistic behavior

Stakeholder-centered typology (D'Arcy and Bender 2023)

	Direct stakeholders	Indirect stakeholders
Tech use	By choice	Harm to community
Tech use	Not by choice	Harm to individual
Tech development	Annotator or crowd worker	Unwitting data contributor

Direct Stakeholder, Tech User, By Choice

- *I choose to use this voice assistant, dictation software, machine translation system…*
    - … but it doesn't work for my language or language variety
        - Suggests that my language/language variety is inadequate
        - Makes the product unusable for me
- … but the system doesn't indicate how reliable it is
    - Users reliant on machine translation/auto-captioning for important info left in the dark about what they might be missing

Direct Stakeholder, Tech User, 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 911 system requires interaction with a virtual agent first*

- … but it doesn't work or doesn't work well for my language variety or speech characteristics (e.g. stutter; Wu 2024)
    - I scored poorly on the interview, even though the content of my answers was good
    - I can't access my account information or 911

Direct Stakeholder, Tech User, Not By Choice

- *LM (language modeling) technology can now generate very real sounding text, in English at least* (Radford et al 2019, Brown et al 2020, Bender, Gebru et al 2021)
    - … but which is not grounded in any actual relationship to facts
        - I mistake the text for statements made by a human publicly committing to them
        - I become more distrustful of all text I see online
- Language models trained on ‘standard’ or ‘official’ sounding documents will sound ‘standard’ or ‘official’.

Direct Stakeholder, Tech Development, Annotator or Crowdworker

- Crowdworking practices are exploitative, often lacking fair compensation, sometimes involving withholding of compensation (Fort et al 2011, Irani & Silberman 2013)

- Subcontracting practices are similarly exploitative, often "benefiting from catastrophe" through out-sourcing to countries suffering economic crises

- The work can be traumatic (e.g. content moderation), and workers often aren't provided sufficient psychological support

- Workers are sometimes drawn from especially vulnerable populations, such as incarcerated people and minors

Indirect Stakeholder, Tech Use, Harm to Individual

- *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)

        - My restaurant's positive reviews are underrated because of the name of the cuisine (Speer 2017)

        - My resume is rejected because the screening system has learned that typically "masculine" hobbies correlate with getting hired

        - My image search reflects stereotypes back to me (Noble 2018)

Indirect Stakeholder, Tech Use, Harm to Individual

- *Annotations provided for supervised learning systems reflect the linguistic competence of the annotators*
    - … which can be poorly matched to the training data & use case
        - Content moderation systems might flag content just for its language variety, because out-group annotators were more likely to label such language as “offensive” (Sap et al 2019)

Indirect Stakeholder, Tech Use, Harm to Individual

- *Someone searched for critics of the government*
    - … and found my blog post/tweet
- *Someone put my words into an MT system*
    - … which got the translation wrong and led the police to arrest me
      (*The Guardian*, 24 Oct 2017; https://bit.ly/2zyEetp)
- *Someone built an identity characteristic classifier*
    - … with a fixed set of identity categories that erases mine
    - … and outed me based on characteristics of my language use

Indirect Stakeholder, Tech Use, Harm to Community

- *Someone searched for me online*
    - … but the search triggered display of negative ads including my name because of stereotypes about my ethnic identity (Sweeney 2013)
- *Virtual assistants are gendered as female and bossed around*

Indirect Stakeholder, Tech Use, Harm to Community

- *LM (language modeling) technology can now generate very real sounding text, in English at least* (Radford et al 2019)
    - … but which is not grounded in any actual relationship to facts
    - Such systems are then used by extremist groups to synthesize text to populate message boards used to radicalize people (McGuffie & Newhouse 2020)
    - Such systems are then used by people seeking profit (ad clicks, book sales) to create non-fiction-looking texts, polluting the information ecosystem (Shah & Bender 2024)

Indirect Stakeholder, Tech Use, Harm to Community

- *Sentiment analysis systems don't work well on my dialect*
    - … my community's input is not included when social media discussions are processed for public policy input
- *Language ID systems don't identify my dialect*
    - … Social-media based disease warning systems fail to work in my community (Jurgens et al 2017)

Indirect Stakeholder, Tech Use, Harm to Community

- *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)
    - autocompletion of search queries repeats & reinforces harmful stereotypes (Noble 2018)
    - chatbots for language learners repeat & reinforce harmful stereotypes
    - chatbots for NPCs in games repeat & reinforce harmful stereotypes

Indirect Stakeholder, Tech Use, Harm to Community

- *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)
        - My restaurant's positive reviews are underrated because of the name of the cuisine (Speer 2017)
        - My resume is rejected because the screening system has learned that typically "masculine" hobbies correlate with getting hired
        - My image search reflects stereotypes back to me (Noble 2018)

Indirect Stakeholder, Tech Use, Harm to Community

- The corporate race for ever larger language models leads to increasing use of energy, water, and rare earth minerals (Strubell et al 2019, Luccioni et al 2023)
    - … leading to environmental degradation for everyone
    - … but more and first for marginalized communities (Bender, Gebru et al 2021)

Stakeholder-centered typology (D'Arcy and Bender 2023)

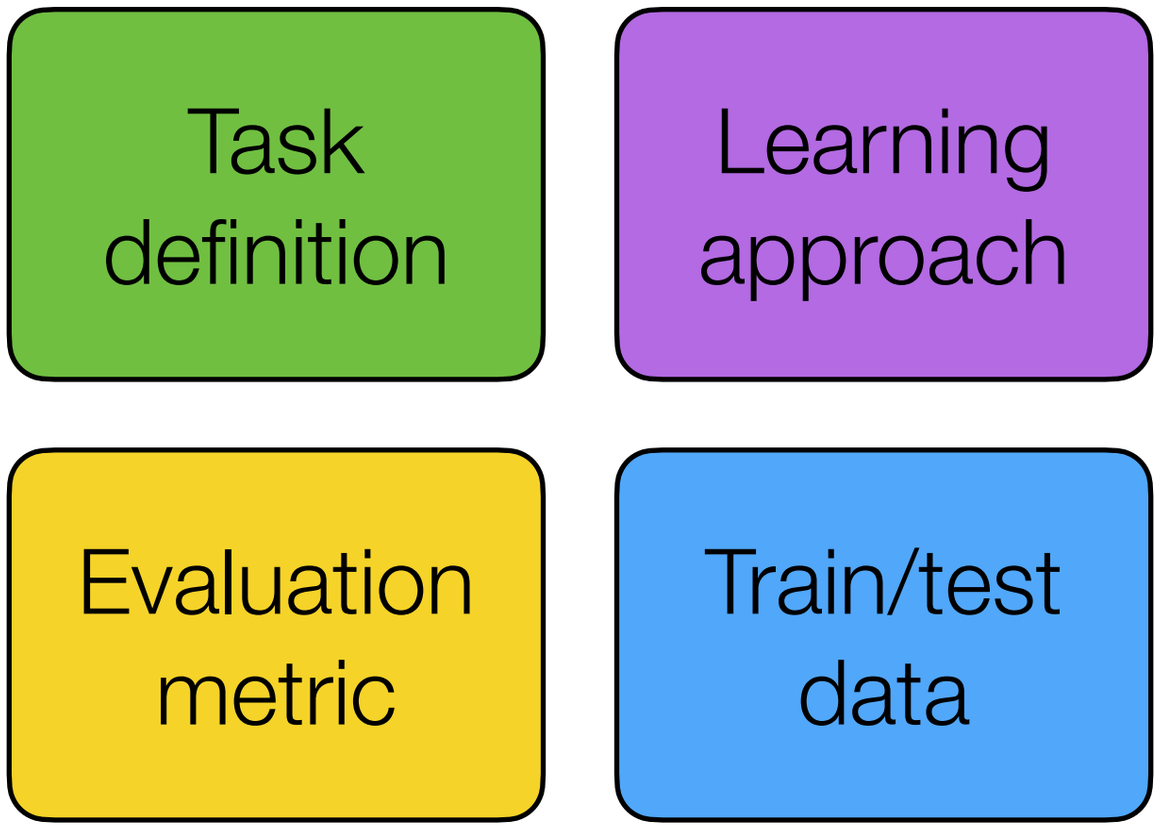
	Direct stakeholders	Indirect stakeholders
Tech use	By choice	Harm to community
Tech use	Not by choice	Harm to individual
Tech development	Annotator or crowd worker	Unwitting data contributor

What does this mean for language technology 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)

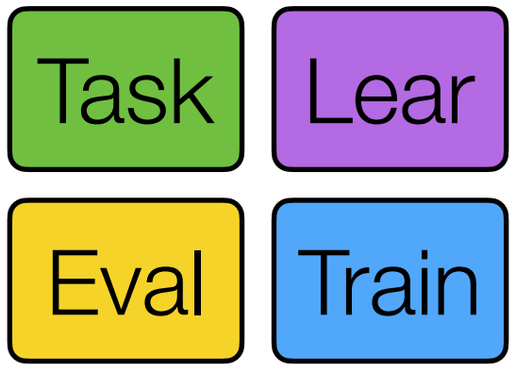


Machine learning, in context

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?

Data Statements for NLP: Transparent documentation

(Bender & Friedman 2018, Bender et al 2021, McMillan-Major et al 2023)

- Foreground characteristics of our datasets (see also: AI Now Institute 2018, Gebru et al 2018, 2021, 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

Data Statements Schema Version 3

(McMillan-Major & Bender, forthcoming)

---

1. Header
2. Executive Summary
3. Curation Rationale
4. Documentation for Source Datasets
5. Language Varieties
6. Language User Demographic
7. Annotator Demographic
8. Linguistic Situation and Text Characteristics
9. Preprocessing and Data Formatting
10. Capture Quality
11. Limitations
12. Metadata
13. Disclosure and Ethical Review
14. Distribution
15. Maintenance
16. Other
17. Glossary

Case: Direct stakeholders whose varieties aren't well represented

- **Speech/language tech researchers & developers:** Map out underrepresented language varieties and direct effort appropriately; test approaches more broadly

- **Procurers:** Is this trained model likely to work for our clientele?

- **Consumers:** Is this trained model likely to work for me?

- **Members of the public:** Advocate for models trained on datasets that are responsive to the community of users

- **Policy makers:** Require automated systems to be *accessible* to speakers of all language varieties in the community

Case: Indirect stakeholders whose varieties aren't well represented

- **Speech/language tech researchers & developers:** Map out underrepresented language varieties and direct effort appropriately; test approaches more broadly

- **Procurers:** What information is this system going to expose and what is it going to miss?

- **Consumers:** Is this software being transparent about how well it can work and under what circumstances it works better/worse?

- **Members of the public:** Advocate for transparency regarding system performance across representative samples

- **Policy makers:** Require broad testing of systems and transparency regarding system confidence/failure modes

Data statements are not a panacea!

- Mitigation of the negative impacts of speech/language technology will require on-going work and engagement (and cost/benefit analysis)

- Data statements are intended as one practice among others that position us (in various roles) to anticipate & mitigate some negative impacts

- Probably won't help with e.g.:
  - impacts of gendering virtual agents
  - privacy concerns around classification of identity characteristics

- Can help with problems stemming from lack of representative data sets and possibly also 'automation bias' (Skitka et al 2000)

Beyond data statements:
What else can we do?

- Consentful data collection, even for “public” data
    - Consent, control, compensation, credit
- Demand and provide transparency into working conditions & pay
- Demand and provide transparency into environmental impact (Henderson et al 2020)
- => Preference for small, purpose-built systems

Beyond data statements: What else can we do?

- 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

Summary

- The L in NLP means language and language means people (Schnoebelen 2017) … and variation!

- When we are working on tech that will be deployed in the world, we need to keep an eye on how it fits into the world

- It's easy to get bogged down in "this is too terrible" or "this is too hard", and then turn away (from NLP or its societal impacts), but we don't have to get stuck there

- Transparency is a good starting point: documentation of datasets & models; clear discussion of application—world relationship; transparency about labor; transparency about environmental impact

Questions for this group

- What are some language technologies you use personally?
- What are some language technologies you are using or developing professionally?
- Which points from this presentation are applicable to those use cases?
- How can the best practices outlined apply?

References

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

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., Freidman, B., and McMillan-Major, A. (2021a). A guide for writing data statements for natural language processing.

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.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021b). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of FAccT 2021*.

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.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

D'Arcy, A. and Bender, E. M. (2023). Ethics in linguistics. *Annual Review of Linguistics*, 9(Volume 9, 2023):49–69.

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

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

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., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):8692.

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.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

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.

Irani, L. C. and Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 611620, New York, NY, USA. Association for Computing Machinery.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (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.

Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 38193828, New York, NY, USA. Association for Computing Machinery.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

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

Lefeuvre-Halftermeyer, A., Govaere, V., Antoine, J.-Y., Allegre, W., Pouplin, S., Departe, J.-P., Slimani, S., and Spagnulo, A. (2016). Typologie des risques pour une analyse éthique de l'impact des technologies du TAL. *Revue TAL: traitement automatique des langues*, 57(2):47–71.

Luccioni, A. S., Jernite, Y., and Strubell, E. (2023). Power hungry processing: Watts driving the cost of AI deployment?

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. Technical report, Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey. `https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf`.

McMillan-Major, A. and Bender, E. M. (forthcoming). A guide for creating and documenting language datasets with data statements schema version 3.

McMillan-Major, A., Bender, E. M., and Friedman, B. (2023). Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.* Just Accepted.

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*, pages 220–229. ACM.

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

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Unpublished MS, OpenAI San Francisco.

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

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.

Shah, C. and Bender, E. M. (2024). Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web*. Just Accepted.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 723741, New York, NY, USA. Association for Computing Machinery.

Skitka, L. J., Mosier, K., and Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4):701 – 717.

Speer, R. (2017). Conceptnet numberbatch 17.04: better, less-stereotyped word vectors.

Blog post, https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/, accessed 6 July 2017.

Strubell, E., Ganesh, A., and McCallum, A. (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.

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

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

Wu, S. (2024). Blocked by the system: How current voice AI silences people who stutter. Paper presented at AAAS 2024, Boulder CO.