A Typology of Ethical Risks in Language Technology with an Eye Towards Where Transparent Documentation Can Help

Emily M. Bender
University of Washington

The Future of AI workshop
University of Cambridge
25 March 2019

Goals

- Present a typology of the risks of adverse impacts of NLP technology
- Present *data statements*: a positive step we can take to position ourselves to mitigate such risks
- Reflect on which types of risks data statements help with
- Reflect on whose job it is to worry about these things

Goals

- Present a typology of the risks of adverse impacts of NLP technology

Non-exhaustive, preliminary

- Present *data statements*: a positive step we can take to position ourselves to mitigate such risks

- Reflect on which types of risks data statements help with

- Reflect on whose job it is to worry about these things

Goals

- Present a typology of the risks of adverse impacts of NLP technology

  Non-exhaustive, preliminary

- Present *data statements*: a positive step we can take to position ourselves to mitigate such risks

  One tool, not a panacea!

- Reflect on which types of risks data statements help with

- Reflect on whose job it is to worry about these things

Goals

- Present a typology of the risks of adverse impacts of NLP technology

Non-exhaustive, preliminary

- Present *data statements*: a positive step we can take to position ourselves to mitigate such risks

One tool, not a panacea!

- Reflect on which types of risks data statements help with

Some, not all

- Reflect on whose job it is to worry about these things

Goals

- Present a typology of the risks of adverse impacts of NLP technology

Non-exhaustive, preliminary

- Present *data statements*: a positive step we can take to position ourselves to mitigate such risks

One tool, not a panacea!

- Reflect on which types of risks data statements help with

Some, not all

- Reflect on whose job it is to worry about these things

Everyone's; in different ways

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, Ovegeneralization, Bias confirmation, Topic Overexposure, Dual use
  - Illustrated with NLP-specific examples of negative impacts
- Not exhaustive, not a typology

The L in NLP is Language, language means people
(Schnoebelen 2017)

- Schnoebelen, summarizing EthNLP 2017 (Hovy et al 2017):
  - Look to NLP (and AI) to assist people, not replace them
  - Engage with scholarly disciplines that have a better understanding of people
- Value sensitive design (Friedman et al 2006, Friedman & Hendry to appear):
  - Identify stakeholders
  - Design to support stakeholders' values

The L in NLP is Language, language means people

Direct stakeholders	Indirect stakeholders
By choice	Subject of query
Not by choice	Contributor to broad corpus
	Subject of stereotypes

Direct stakeholders: By choice

- *I choose to use this spell checker, autocorrect, voice assistant, MT 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 MT/auto-captioning for important info left in the dark about what they might be missing

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*
    - … 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

Direct stakeholders: Not by choice

- *LM 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
        - I mistake the text for statements made by a human publicly committing to them
        - I become more distrustful of all text I see online

Indirect stakeholders: Subject of query

- *Someone searched for me online*
  - … but the search triggered display of negative ads including my name because stereotypes about my ethnic identity (Sweeney 2013)
- *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)

Indirect stakeholders: Subject of query

- *Sor*
  - ative
- *Sor*
- *Sor*

Facebook

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

Indirect stakeholders: Subject of query

- *Someone designed a system to classify people by identity characteristics according to linguistic features*

    - Information I thought I was presenting only in some venues is made available in others

Indirect stakeholders: Contributor to broad corpus

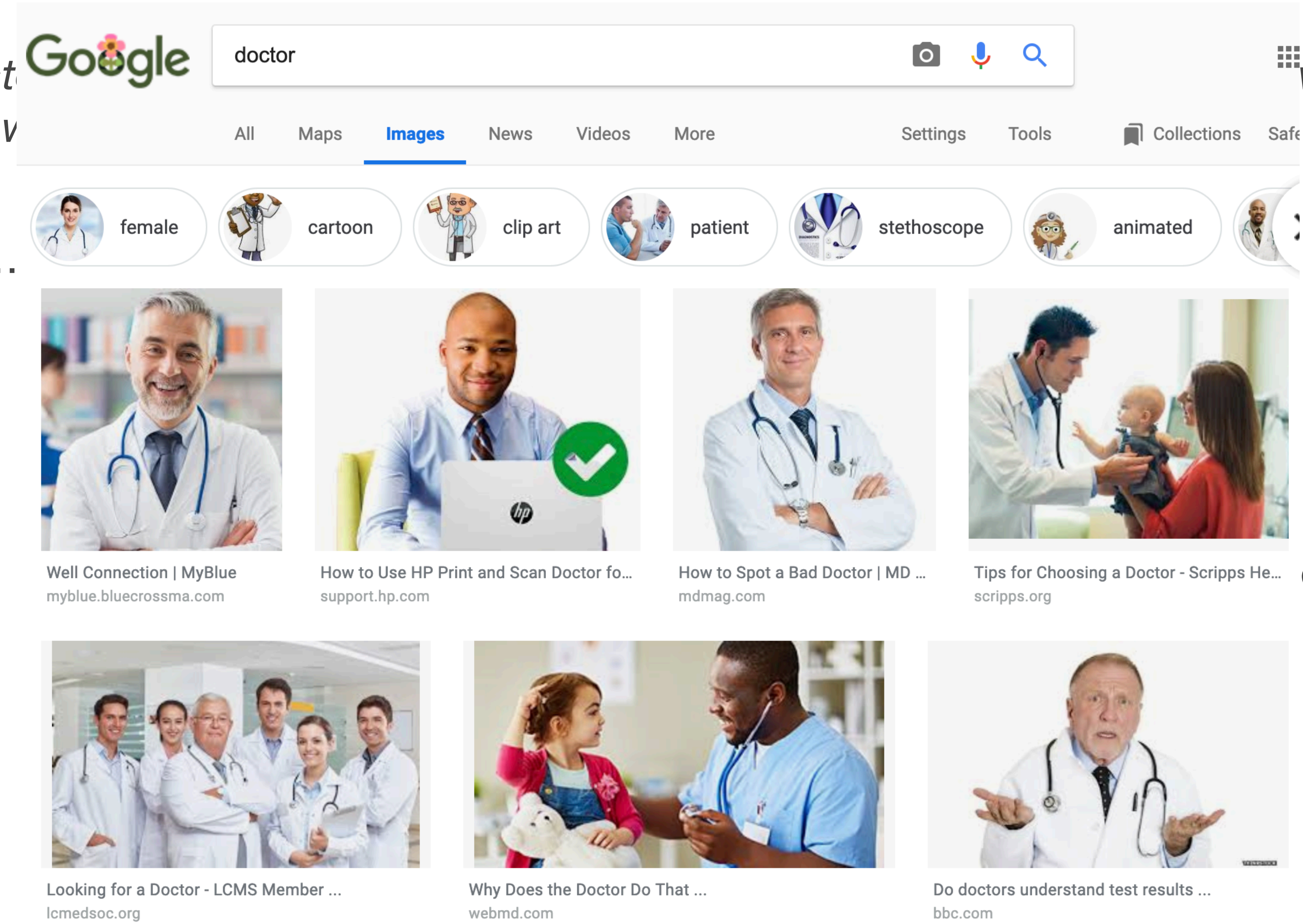
- *ASR doesn't caption my words as well as others'*
    - My contributions are rendered invisible to search engines
- *Language ID systems don't identify my dialect*
    - Social-media based disease warning systems fail to work in my community (Jurgens et al 2017)

Indirect stakeholders: Subject of stereotypes

- *Virtual assistants are gendered as female and ordered around*
- *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

Indirect stakeholders: Subject of stereotypes

- *Syst* *world*
  knov
  name
  ed
  - ..



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

Proposed Schema: Long Form

- A. Curation Rationale
- B. Language Variety
- C. Speaker Demographic
- D. Annotator Demographic
- E. Speech Situation
- F. Text Characteristics
- G. Recording Quality
- H. Other
- I. Provenance Appendix

What kind of language behavior?

Proposed Schema: Long Form

What data? Why?

- A. Curation Rationale
- B. Language Variety
- C. Speaker Demographic
- D. Annotator Demographic
- E. Speech Situation
- F. Text Characteristics
- G. Recording Quality
- H. Other
- I. Provenance Appendix

What kind of language behavior?

Proposed Schema: Long Form

- A. Curation Rationale
- B. Language Variety
- C. Speaker Demographic
- D. Annotator Demographic
- E. Speech Situation
- F. Text Characteristics
- G. Recording Quality
- H. Other
- I. Provenance Appendix

What data? Why?

Whose language?

What kind of language behavior?

Proposed Schema: Short Form

- 60-100 word summary of the information in long form data statement, hitting most main points
- Include pointer to where the long form can be found
- Paper presenting the dataset originally
- Project web page
- System documentation

Who’s job is this?

- **NLP researchers & developers:** build better systems, promote systems appropriately, educate the public
- **Procurers:** choose systems/training data that match use case, align task assigned to NLP system with goals
- **Consumers:** understand NLP output as the result of pattern recognition, trained on some dataset somewhere
- **Members of the public:** learn about benefits and impacts of NLP and advocate for appropriate policy
- **Policy makers:** consider impacts of pattern matching on progress towards equity, require disclosure of characteristics of training data

Case: Direct stakeholders whose varieties aren’t well represented

- **NLP 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 subject to stereotypes

- **NLP researchers & developers:** Conceptualize training text as things specific people have said, rather than unproblematic ‘common sense knowledge’
- **Procurers:** What kind of text underlies the system I’m purchasing and how do the tasks I’m setting for it risk amplifying biases from that text?
- **Consumers:** Know what is the ultimate source of this information I’m seeing and understand it as the viewpoints of people (aggregated)
- **Members of the public:** Advocate for transparency
- **Policy makers:** Require automated systems to be *transparent* about sources of ‘knowledge’

Data statements are not a panacea!

- Mitigation of the negative impacts of NLP 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

But they may help in combating automation bias
(Skitka et al 2000)

- By foregrounding characteristics of training datasets, foreground:
  - The L in NLP means Language, language means people
  - The datasets NLP systems are trained on ultimately come from people, speaking about certain topics, for a certain purpose
- Treat text-derived ‘common sense’ with skepticism, understand where it is being used
- Understand machine output as pattern matching against specific (if large) datasets, not expert decision making

Lessons from sociolinguistics
(e.g. Labov 1966, Eckert & Rickford 2001)

- Variation is the natural state of language
- Meaning, including social meaning, is negotiated in language use
- Our social world is largely constructed through linguistic behavior

- Keeping these lessons in focus will help us make better, more responsive natural language technology

Lessons from sociolinguistics
(e.g. Labov 1966, Eckert & Rickford 2001)

- Variation is the natural state of language
- Meaning, including social meaning, is negotiated in language use
- Our social world is largely constructed through linguistic behavior

- Keeping these lessons in focus will help us make better, more responsive natural language technology

Thank you!

References

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

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.

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 D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. 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.

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

Friedman, B. and Hendry, D. (To appear). *Value Sensitive Design: A twenty-year synthesis and retrospective*.

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

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2018). 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. arXiv:1903.03862v1.

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.

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.

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.

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

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.

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.

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.

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.

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