Yulia Tsvetkov's
Computational Ethics in NLP, Winter 2022
Reading List

1. **1/11/2022**:Introduction:Motivation, course overview and requirements. Examples of projects in computational ethics
    - Hovy & Spruit (2016) The Social Impact of NLP. ACL.
    - Barocas & Selbst (2016) Big Data's Disparate Impact. California Law Review 671
    - Barbara Grosz talk (2017) Intelligent Systems: Design & Ethical Challenges
    - Kate Crawford NeurIPS keynote (2017) The Trouble with Bias
    - Yonatan Zunger blog post (2017) Asking the Right Questions About AI

2. **1/13/2022**:Human subjects research:History: medical, psychological experiments, IRB and human subjects. Participants, labelers, and data in NLP
    - The Belmont Report
    - The Menlo Report (Ethical Principles Guiding Information and Communication Technology Research)
    - Williams, Burnap & Sloan (2017) Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.
    - Vitak, Shilton & Ashktorab (2016) Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. CSCW.
    - Crowdsourcing
        - Many more relevant readings in Chris Callison-Burch's class http://crowdsourcing-class.org/

3. **1/18/2022**:Human subjects research: **Paper discussions**
    - Group 1: Matthew L Williams, Pete Burnap & Luke Sloan (2017) Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.
    - Group 2: Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. (2020) Datasheets for Datasets. Arxiv.
    - Group 3: Mor Geva, Yoav Goldberg, Jonathan Berant (2019) Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets EMNLP
    - Group 4: R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? ACM FAT*

4. **1/20/2022**:Philosophical foundations:Ethical frameworks, benefit and harm, power, automation. **Optional readings:**

- AI Ethics: Seven Traps. Annette Zimmermann and Bendert Zevenbergen, 2019.
- The Values Encoded in Machine Learning Research. Birhane *et al.*, 2021.
- The Steep Cost of Capture. Meredith Whittaker, *Interactions*, 2021.
- Negishi Welfare Weights: The Mathematics of Inequality. Elizabeth Stanton, 2009.
  - Negishi weighting (e.g., of regional utility functions) has been used in the design of macroeconomic policy and treaties, including (according to several sources I found, though none seemed authoritative) the Kyoto Protocol. According to Stanton, it is often portrayed as a "standard technical assumption," but actually is extremely value-laden. More on the dispute is on Wikipedia, and a response to her criticism is here.
- Physiognomy's New Clothes. Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, 2017.
- The Ethics of Belief. William K. Clifford, *Contemporary Review*, 1877.
- [pending more]

5. **1/25/2022**:Philosophical foundations: **Paper discussions**
   - Group 1: Race After Technology. Ruha Benjamin, 2019.
     - Introduction only.
     - Benjamin provides examples of the power of language and categories (e.g., the power of names, and the messages that are sent by them). How do these categories interact with existing power structures in society, and how can they be used to subvert those power structures? What role might language technology play in this?
   - Group 2: The Foundations of Bioethics, 2nd Ed. H. Tristram Engelhardt, 1996.
     - Read Ch. 1 (pp. 3–16) and Ch. 3 (pp. 121–131 only, beginning with "the principles of bioethics").
     - Engelhardt presents a deontology designed to deal with the problem of a morality between irreconcilable "moral strangers." He highlights the tension between permission/autonomy and beneficence/welfare. How does the role of the doctor in his ethics compare to the role of the technologist or researcher? Do his arguments translate over?
   - Group 3: Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. John C. Harsanyi, *Journal of Political Economy,* 1955.
     - Skip the proofs and read the informal descriptions of the postulates/theorems, which are fairly clear.
     - In a possible contrast to the contra-Delphi arguments about ethics not being an "average", Harsanyi explicitly argues for estimating average utility as an ideal social welfare function. This also contrasts with the Veil of Ignorance of Rawls, who looks at the welfare of the *worst*-off agent. Do you buy Harsanyi's argument? Under what conditions would you agree that maximizing a Harsanyi-style social welfare function is ideal? What assumptions of his do you take issue with, if any?

- Group 4: Delphi: Towards Machine Ethics and Norms. Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi, *ArXiv*, 2021.
    - See the authors' follow-up post for more detail on the goals of the research and clarification of points in the paper.
    - How do the arguments in the paper and post fit into the broader milieu of ethical philosophy? What use might a preference utilitarian, permission-driven deontologist, or social contract theorist have for a model like Delphi?

6. **1/27/2022**:Social bias and algorithmic (un)fairness:Psychological foundations of bias; social bias and disparities in NLP data and models.
    - NIPS Keynote: Kate Crawford, The Trouble with Bias
    - Implicit bias: scientific foundations
        - Stereotypes and Prejudice: Their Automatic and Controlled Components
        - Stereotype threat
        - Breaking the prejudice habit: Mechanisms, timecourse, and longevity.
        - The evolution of cognitive bias
        - Psychological experiments to quantify bias: IAT ; (Greenwald et al. 1998)
    - Summary of non-computational work on microaggressions and the effect of biased attitudes on minorities and marginalized groups: Microaggressions towards racial/ethnic groups [ref1, ref2,]; Qualitative research involving African Americans (Sue, Capodilupo, & Holder, 2008), Asian Americans (Sue, Bucceri, Lin, Nadal, & Torino, 2010), and Latina/o Americans (Rivera, Forquer, & Rangel, 2010) have supported that members of these groups experience microaggressions in their everyday lives and that such experiences have a negative toll on psychological well-being; Microaggressions towards gender [ref1, ref2, ref3, ref4]; Microaggressions towards sexual orientation [ref]; Microaggressions towards physical disability [ref]; Microaggressions towards persons with mental illness [ref]; Microaggressions towards people from religious minority groups: Muslim Americans [ref]; Incivility breeds more incivility [Foulk et al. 2014]; Incivility leads to stress, depression, and lack of commitment [Miner et al. 2012, Lim et al. 2008]; Women are more likely to experience incivility [Cortina et al. 2008]; our work on condescension [TODO: add a pointer];
    - Gender bias in the job market: a longitudinal analysis [Tang et al. 2017]
    - On gender bias in tech jobs [ref]
        - In open-source software development [Vedres & Vasarhelyi 2019]
        - In NLP [Schluter 2018]
    - TechCrunch: 5 unexpected sources of bias in AI
    - Daumé III's blog post (2016) Bias in ML, and Teaching AI
    - Biases revealed in online data or downstream applications
        - The dominant class is often portrayed and perceived as relatively more professional (Kay, Matuszek, and Munson 2015)

        - Males are over-represented in the reporting of web-based news articles (Jia, Lansdall-Welfare, and Cristianini 2015)
        - Males are over-represented in twitter conversations (Garcia, Weber, and Garimella 2014)
        - Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues (Wagner et al. 2015)
        - IMDB reviews written by women are perceived as less useful (Otterbacher 2013)
        - Biases in in the Flickr30K Dataset (Miltenburg 2016)
        - Gender and Dialect Bias in YouTube's Automatic Captions (Tatman 2017)
        - Social Bias in Elicited Natural Language Inferences (Rudinger, May, Van Durme 2017)
        - Racial disparities in off-the-shelf LID tools (Blodgett & O'connor 2016)
        - Bias in facial recognition: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html
        - Gender bias in syntactic n-grams corpus (Hoyle et al. ACl 2019)
        - Model bias in detecting aggression in on Twitter (Zhong et al. EMNLP 2019)
    - Bias in facial recognition systems: (Buolamwini and Gerbu, 2018, FAT*)
    - Quantifying biases: integrating statistical models with (socio)linguistic theories
        - Respect: Voight et al. (2017) Language from police body camera footage shows racial disparities in officer respect, PNAS
        - Affect Control Theory: Joseph et al. (2017) Girls rule, boys drool: Extracting semantic and affective stereotypes on Twitter, CSCW
    - Quantifying biases in text corpora (organized by key approach in the paper)
        - Lexicon-based approaches
            - + Regression/classification (Voight et al. '17 [race]; Recasens et al. '13)
            - + Crowdsourcing (Fast et al. '16) [gender]
        - Language models (Fu et al. '16) [gender]
        - Sociolinguistic knowledge
            - +Respect (Voight et al. '17)
            - +ACT+Latent variable models (Joseph et al. '17) [gender]
        - Word embeddings (Bolukbasi et al. '16 [gender]; Caliskan et al. '17; Manzini et al. '19 [race, religion]; Kurita et al.'19, Zhao et al.'19 [contextualized embeddings])
        - Measuring bias amplification in trained models (Zhao et al. '17) [gender]
        - Latent variable models, gender bias in syntactic n-grams corpus (Hoyle et al. ACl 2019)
    - Measuring bias in natural language generation models (Sheng et al. EMNLP 2019)
    - Debiasing

        - + Transforming embeddings: Bolukbasi et al. (2016) Debiasing word embeddings, NIPS
        - + Regularization: Zhao et al. (2017) Reducing gender amplifications in models, EMNLP
        - + Resampling training data: Jurgens et al. (2017) Incorporating Dialectal Variability for Socially Equitable Language Identification [socioeconomic status]
        - Data balancing for morphologically rich languages (Zmigrod et al. ACL 2019)

    - Identifying social stereotypes (in social media or in fictional worlds (novels, movies), to understand how media shapes and reflects social perceptions)
        - Fast et al. (2016) Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community, ICWSM
        - Sap et al. (2017) Connotation Frames of Power and Agency in Modern Films, EMNLP
        - Bamman et al. (2014) A Bayesian Mixed Effects Model of Literary Character, ACL
        - Beller et al (2013) I'm a Belieber: Social Roles via Self-identification and Conceptual Attributes, ACL
        - Carpenter et al.(2016) Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes , SPPS
        - Flekova et al. (2016) Analyzing Biases in Human Perception of User Age and Gender from Text, ACL
    - Effects of annotator biases on crowd-sourced annotations
        - In NLU data sets (Geva et al. EMNLP 2019)
        - In hate speech data sets (Sap et al.. ACL 2019, Davidson et al. Abusive Language Workshop at ACL 2019)
    - Surveys
        - Gender bias studies in NLP: Sun et al. (2019) Mitigating Gender Bias in Natural Language Processing: Literature Review *ACL*
        - Racial bias studies in NLP: Field et al. (2021) A Survey of Race, Racism, and Anti-Racism in NLP *ACL*

7. **2/1/2022**:Social bias in NLP models: **Paper discussions**
    - Group 1: Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. (2021) Intrinsic bias metrics do not correlate with application bias. ACL.
    - Group 2: Tomalin, Marcus, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. (2021) The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*.
    - Group 3: Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach.

(2020) [Language (Technology) is Power: A Critical Survey of "Bias" in NLP](). ACL
- Group 4: Nangia, Nikita, Clara Vania, Rasika Bhalerao, and Samuel Bowman. (2020) CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. EMNLP

8. **2/3/2022**:NLP for detecting bias and stereotypes: **Paper discussions**
    - Group 1: Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. "Language from police body camera footage shows racial disparities in officer respect." Proceedings of the National Academy of Sciences 114, no. 25 (2017): 6521-6526.
    - Group 2: Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 596-608. 2020.
    - Group 3: Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. "Social Bias Frames: Reasoning about Social and Power Implications of Language." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5477-5490. 2020.
    - Group 4: Fraser, Kathleen C., Isar Nejadgholi, and Svetlana Kiritchenko. "Understanding and Countering Stereotypes: A Computational Approach to the Stereotype Content Model." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL). 2021.

9. **2/8/2022**:Hate speech:NLP for identifying and countering hate speech/toxicity/abuse
    - Surveys, books, book chapters
        - Definition of Hate Speech (Nockleby, J. *Encyclopedia of the American Constitution* 2000)
        - Equality and Freedom of Expression: The Hate Speech Dilemma (Toni M. Massaro, William & Mary Law Review 1990)
        - A Survey on Hate Speech Detection using Natural Language Processing (Schmidt & Wiegand SocialNLP'17)
        - Understanding Abuse: A Typology of Abusive Language Detection Subtasks (Waseem et al. 1st Workshop on Abusive Language Online'17)
        - Gendered Cyberhate, Victim-Blaming, and Why the Internet is More Like Driving a Car on a Road Than Being Naked in the Snow
        - Misogyny Online
        - F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti (2021) Resources and benchmark corpora for hate speech detection: a systematic review. In Lang Resources & Evaluation
        - B. Vidgen & L. Derczynski (2020) Directions in Abusive Language Training Data: Garbage In, Garbage Out PLOS One
        - Wenjie Yin & Arkaitz Zubiaga (2021) Towards generalisable hate speech detection: a review on obstacles and solutions In *PeerJ Comput Sci*

        - Kiritchenko, Svetlana, Isar Nejadgholi, and Kathleen C. Fraser. "Confronting abusive language online: A survey from the ethical and human rights perspective." Journal of Artificial Intelligence Research 71 (2021): 431-478.
    - Workshops
        - 1st Workshop on Abusive Language Online (WOAH 2017)
        - WOAH 2 (2018)
        - WOAH 3 (2019)
        - WOAH 4 (2020) https://aclanthology.org/volumes/2020.alw-1/
        - WOAH 5 (2021) https://aclanthology.org/events/woah-2021/
        - WOAH 6 https://www.workshoponline abuse.com/
        - Gendered violence online: a scholarly 'slam'
    - Computational approaches to detecting hate speech, abusive and toxic language
        - Detecting Hate Speech on the World Wide Web (Warner & Hirschberg LSM'12)
        - Abusive Language Detection in Online User Content (Nobata et al. WWW'16)
        - Analyzing the Targets of Hate in Online Social Media (Silva et al. ICWSM'16)
        - Hate Speech Detection with Comment Embeddings (Djuric et al. WWW'15)
        - Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter (a good example of the offline effects of online abuse Chatzakou et al. *ACM Hypertext*'17)
        - The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data (Chandrasekharan et al. CHI'17)
        - Automated Hate Speech Detection and the Problem of Offensive Language (Davidson et al. ICWSM'17)
        - Mean Birds: Detecting Aggression and Bullying on Twitter (Chatzakou et al. arxiv'17)
        - Using Convolutional Neural Networks to Classify Hate-Speech (Gambäck & Sikdar 1st Wrshp on Abusive Language'17)
        - A Unified Deep Learning Architecture for Abuse Detection (Founta et al. AAAI'18)
        - Google Perspective API https://arxiv.org/pdf/1702.08138.pdf
        - Ethical Challenges in Data-Driven Dialogue Systems (Henderson et al. *arxiv*'17)
        - Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter (Waseem and Hovy, NAACL 2016)
        - The Linguistic Ideologies of Deep Abusive Language Classification (Castelle, ALW '18)
        - The Risk of Racial Bias in Hate Speech Detection (Sap et al., ACL 2019)
        - Effects of moderation of hate speech: (Chandrasekharan et al., CHI 2017)
    - Other types of toxic behaviors in communication

        - Trolling (Cheng et al. ICWSM'15, CSCW'17)
        - Politically incorrect language: (Emile Hine et al. ICWSM 2017)
    - Fighting hate speech with counterspeech
        - Susan Benesch et al. (2016) Counterspeech on Twitter: A Field Study.
        - Mathew B. et al. (2019) Thou Shalt Not Hate: Countering Online Hate Speech. ICWSM
        - Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In Proceedings of ACL.
        - Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In Proceedings of ACL
        - Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding and William Yang Wang, (2019) A Benchmark Dataset for Learning to Intervene in Online Hate Speech EMNLP
    - Leaderboards:
        - https://paperswithcode.com/task/hate-speech-detection
    - Tutorial plans:
        - https://dl.acm.org/doi/pdf/10.1145/3488560.3501392
        -

10. 2/10/2022:Hate speech: **Paper discussions**
    - Group 1: Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. (2021) Challenges in Automated Debiasing for Toxic Language Detection. In Proc. EACL
    - Group 2: Paul Rottger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert (2021) HATECHECK: Functional Tests for Hate Speech Detection Models In *Proc. ACL*
    - Group 3: Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury and Diyi Yang (2021) Latent Hatred: A Benchmark for Understanding Implicit Hate Speech In *Proc. EMNLP*
    - Group 4: Xiaochuang Han, Yulia Tsvetkov (2020) Fortifying Toxic Speech Detectors Against Veiled Toxicity In *Proc. EMNLP*

11. 2/15/2022:Misinformation:NLP for fact-checking and fake news detection. Computational propaganda and political misinformation
    - Fake news/Fact verification
        - Method for real-time fact checking, including a discussion of major challenges (Computation+Journalism 2019)
        - Discussion of ClaimReview Markup and method for retrieving documents that are relevant to human-generated fact-checks (WWW 2018)
        - Fact Extraction and Verification Workshop (FEVER)

        - Exploration and mitigation of biases in the FEVER data set (EMNLP 2019)
        - Adversarial attacks on fact-checking systems (EMNLP, 2019)
        - Survey of field across disciplines, highlighting the importance of evidence (COLING 2018)
        - r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection (Nakamura et al. LREC 2020)
        - CoVerifi: A COVID-19 news verification system (Kolluri & Murthy In Online Social Networks and Media)
        - Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News (Vo & Lee EMNLP 2020)
        - A survey on stance detection for mis-and disinformation identification (Hardalov et al. arxiv)
    - Propaganda and media bias (newspaper articles)
        - “In Plain Sight”, annotated data set for types of media bias (Fan et al. EMNLP 2019)
        - Shared task and data set on detecting propaganda strategies in newspaper articles (NLP4IF Workshop 2019 and SemEval 2020 Shared Task)
        - Media manipulation strategies in Russian Newspaper articles (Field et al. EMNLP 2018)
        - BREAKING! Presenting Fake News Corpus for Automated Fact Checking (Pathak & Srihari ACL 2019)
        - Fine-Grained Analysis of Propaganda in News Articles (San Martino et al. EMNLP 2019)
    - Propaganda/manipulation on social media
        - “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument (King et al. American Political Science Review 2017)
        - Starbird, K., & Wilson, T. (2020). Cross-platform disinformation campaigns: Lessons learned and next steps. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-002
    - Censorship
        - Eddie Yang and Margaret E. Roberts (2021) Censorship of Online Encyclopedias: Implications for NLP Models In *Proc. FAccT*
    - Neural misinformation
        - Defending Against Neural Fake News (Zellers et al. NeurIPS 2019)
        - A Decade of Social Bot Detection
        - The Limitations of Stylometry for Detecting Machine-Generated Fake News (Schuster et al. Computational Linguistics)
        - Approaches to evaluating the factual consistency of generated texts ()

12. 2/17/2022:Misinformation: **Paper discussions**

- Group 1: Eddie Yang and Margaret E. Roberts (2021) Censorship of Online Encyclopedias: Implications for NLP Models In *Proc. FAccT*
- Group 2: Yuanzhi Chen and Mohammad Rashedul Hasan (2021) Navigating the Kaleidoscope of COVID-19 Misinformation Using Deep Learning In *Proc. EMNLP*
- Group 3: Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch and Dan Roth (2019) Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims In *Proc. NAACL*
- Group 4: Tal Schuster, Roei Schuster, Darsh J Shah and Regina Barzilay (2020) The Limitations of Stylometry for Detecting Machine-Generated Fake News *Computational Linguistics*

13. **2/22/2022**:Privacy:Privacy and anonymity in NLP. Writer profiling and adversarial defenses.
    - Advertising and Microtargeting
        - Investigating sources of PII used in Facebook's targeted advertising (Proceedings on Privacy Enhancing Technologies, 2019)
        - "Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest" (Best Paper award CSCW 2018)
        - TheGuardian article on Cambridge Analytica
        - User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection (IEEE Computer Society 2018)
    - Web tracking / Author profiling / Deanonymization
        - On the dual use of profiling techniques: How Despots Use Twitter to Hunt Dissidents
        - "The Princeton Web Transparency and Accountability Project" (Transparent Data Mining for Big and Small Data, 2017)
        - Discriminating Gender on Twitter. (EMNLP 2011)
        - How well can machine learning predict demographics of social media users? (2017)
        - Writer Profiling Without the Writer's Text SocInfo'17
        - On the Feasibility of Internet-Scale Author Identification (Text-based via writing style, IEEE Symposium on Security and Privacy, 2012)
        - Personal Information Leakage Detection in Conversations. EMNLP'20
        - Broad overview (Manuscript, 2019)
    - Obfuscation / Privacy protection
        - Obfuscating Document Stylometry to Preserve Author Anonymity (ACL 2006)
        - Obfuscating Gender in Social Media Writing (CSS+NLP 2016)
        - Deep Reinforcement Learning-based Text Anonymization against Private-Attribute Inference (EMNLP 2019)
        - Towards Differentially Private Text Representations (SIGIR'20)
        - TextHide: Tackling Data Privacy in Language Understanding Tasks (Findings of EMNLP 2020)

        - [Large language models can be strong differentially private learners](#) (arxiv'21)
    - Using NLP to improve user understanding of privacy
        - Q&A system for privacy policies ([EMNLP 2019](#))
    - Slides from previous talks on Privacy+NLP
        - Lectures at CMU [1](#), [2](#), [3](#)
        - Lecture at UW [1](#)
    -

14. **2/24/2022**:Privacy: **Paper discussion**
    - Group 1: Huang, Yangsibo, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. (2020) "*TextHide*: Tackling data privacy in language understanding tasks." In *Proc. Findings of EMNLP*
    - Group 2: Mireshghallah, Fatemehsadat, Huseyin A. Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. (2021) "Privacy regularization: Joint privacy-utility optimization in language models." In *Proc. NAACL*
    - Group 3: Xu, Qiongkai, Lizhen Qu, Zeyu Gao, and Gholamreza Haffari. (2020) "Personal Information Leakage Detection in Conversations." In *Proc. EMNLP*
    - Group 4: Coavoux, Maximin, Shashi Narayan, and Shay B. Cohen. (2018) "Privacy-preserving neural representations of text." In *Proc. EMNLP*

15. **3/1/2022**:Green NLP: **Paper discussions**
    - Group 1+3 Peter Henderson, Lauren Gillespie, Dan Jurafsky (2021) Environment (Sec. 5.3 in On the Opportunities and Risks of Foundation Models, pp. 139–144)
    - Group 2+4 Emma Strubell, Ananya Ganesh, Andrew McCallum (2019) "Energy and Policy Considerations for Deep Learning in NLP" In *Proc. ACL*

Additional readings on Green NLP:

- Kadan Lottick, Silvia Susai, Sorelle A. Friedler, Jonathan P. Wilson (2019) Energy Usage Reports: Environmental awareness as part of algorithmic accountability In *Proc. NeurIPS*
- Roy Schwartz, Jesse Dodge, Noah A. Smith, Oren Etzioni (2020) Green AI Communications of the ACM
- Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, Joelle Pineau Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning *Preprint*
- CodeCarbon
- Mostafa Dehghani, Anurag Arnab, Lucas Beyer, Ashish Vaswani, Yi Tay (2021) The Efficiency Misnomer *Preprint*
- Lynn Kaack, Priya Donti, Emma Strubell, George Kamiya, Felix Creutzig, David Rolnick (2022) Aligning artificial intelligence with climate change mitigation *Preprint*