Ethics and Power in NLP

Katy Felkner, based on notes by Jonathan May

September 27, 2023

There are entire courses on ethics in AI and NLP. A list of them is here: `https://aclweb.org/aclwiki/Ethics_in_NLP`. Obviously, this lecture won't cover everything relevant, or even important. **The views in these lecture notes are mostly Katy's and occasionally Jon's.**

[3] is a recent critical survey of self-described "bias" papers in NLP. One of the main issues they found was a lack of agreed-upon definitions for terms like "ethics," "harm," and "bias." We're going to spend a significant portion of this lecture discussing and agreeing upon those definitions, so that everyone leaves well-equipped to critically assess papers in this area and their application to your own work.

We'll be doing discussions in groups of 3 today, so form groups now! 2 or 4 are fine, as necessary.

1 Defining "Ethics"

1.1 Philosophical Definitions

1.1.1 Deontological Ethics

- Actions themselves have an inherent, context-independent moral value
- There is a universal set of moral laws, and these laws determine whether an action is right or wrong
- These laws might be based on universal laws of nature, divine commands, or cultural values (depends on which philosopher you ask)
- Morality of an action depends on the actor's *intent* to follow these laws
- "Nothing in the world ... could be called good without qualification except a good will," Immanuel Kant, *Groundwork of the Metaphysic of Morals*, 1785.
- "Act only according to that maxim by which you can also will that it would become a universal law," Kant, *ibid.*
- Criticisms: No law can possibly be universal, so some actions will always be misclassified. Rigid application of absolute laws leads to undesirable consequences.

1.1.2 Consequentialist Ethics

- Moral value of an action is not independent of context.
- Morality of an action should be judged based on its outcome/consequences (or a set of rules based on outcomes).
- Utilitarianism - act to ensure the most good for the most people.
- “It is the greatest happiness of the greatest number that is the measure of right and wrong,” Jeremy Bentham, *A Fragment on Government*, 1776.
- “Logic dictates that the needs of the many outweigh the needs of the few,” Spock, *Star Trek II: The Wrath of Khan*, 1982.
- Criticisms: No obvious way to define what is ”best” for everyone in every way. Can still lead to undesirable individual outcomes. Also, assumes that people in power know what is best for others and will act in their interests.
- Egoism - everyone should act in their own self-interest.
- Criticisms: Assumes that everyone is equally able to act in their own interests, when this is not really the case.

1.2 Discussion

In your group of 3, discuss one or more of these questions.

- What influences have informed your perspective on ethics? Family? Culture? Identity? Education? Religion? Life experiences?
- Do you agree with either the deontological or the consequentialist philosophy? Why or why not?
- What influences should or should not be important in defining a community’s shared ethical principles?

2 Defining “Harm”

2.1 Classifications and Taxonomies of Harms

A recently published taxonomy of language model harms [21] identified 21 harms in 6 categories:

1. **Discrimination, Hate, Exclusion:**
    - social stereotypes and related discrimination - direct emotional harm to stereotyped groups, misleading and incorrect information for others, unfair and inaccurate associations between groups and undesirable concepts

- offensive or hateful speech - even toxicity classifiers can be biased e.g. with high false positive rate on African-American English [17].
- exclusionary norms - related to social stereotypes. LM has incorrect assumptions about who can do what jobs, definitions of marriage and family, etc. Psychological harm to affected groups and potential downstream harms depending on model deployment. Also includes concerns about frozenness and lack of mutability in LMs - language changes, and models need to be able to cope with that.
- performance disparities across social groups - not everyone can benefit equally from a technology, even if everyone is equally at risk of harm
- poor performance on or total exclusion of most languages - NLP systems are often profit-driven, so data and models tend to exist for languages that are likely to be large, profitable markets (e.g. English, Mandarin). Less widely-spoken languages have poorer data coverage and therefore fewer and lower-quality models. Indigenous languages and developing regions are especially underserved.

2. **Information Hazards:**
   - leaking private information memorized from training data. [5] was able to extract physical addresses of real people (who were not public figures) from GPT-2
   - inferring (correctly or incorrectly) sensitive or private information - can build detailed psychological models of individuals without their knowledge or consent. This was possible pre-LLM, and it's likely going to get even easier with LLMs. For example: Facebook-Cambridge Analytica scandal.
3. **Misinformation and factual errors (without malicious user intent):**
   - LM produces false or misleading info
   - LM causes material harm by producing false info in a sensitive domain, e.g. dangerous medication dosages or incorrect legal advice
4. **Malicious Uses:**
   - cheap and easy disinformation campaigns - influence elections, stock prices, crypto markets, etc.
   - easy code generation for malware - now you don't even need to know how to code?
   - facilitating fraud and scams - e.g. extremely convincing personalized phishing emails
   - "illegitimate" surveillance and censorship - spy on millions with relatively few human analysts
5. **HCI Harms/Conversational Agents:**
   - promoting stereotypes with implied gender/ethnicity of agent - Alexa, Siri, Google Assistant use female voices by default, which (arguably) reinforces the stereotype that women should be subservient

- anthropomorphization of systems that seem human - overreliance and excessive trust, user's willingness to share private information, potential for agent to manipulate or influence user

6. **Environmental and Socioeconomic Harms:**

   - environmental issues from LM operation - energy costs for training (significant, per [19]) and inference (possibly greater than training, but much less studied [15])
   - exacerbating income inequality - jobs lost to automation and replaced with relatively few high-paying jobs (e.g. engineering, research) and mostly lower-paying "last-mile" jobs (e.g. content moderation)
   - undermining creative economy with AI-generated content, even in the style of specific writers or artists. Example: WGA strike that just ended had several AI provisions in their tentative agreement (TA), including: AI cannot be used to write or rewrite TV/movie scripts, AI-generated content cannot be used to undermine the right to writing credit, writers cannot be forced to use AI tools in their work, and companies must disclose if materials given to writers are wholly or partially AI-generated.
   - disparate allocation of LM benefits and harms - some groups (usually, those who are already at an advantage) gain the most benefit from LMs, while other groups (usually those who are historically at a disadvantage) stand the greatest risk of harm. This is deeply related to the above discussion of discrimination, hate, and exclusion.

Even in such a comprehensive taxonomy, there are still potential harms of AI/NLP/LMs excluded. The authors note the following exclusions: working conditions of data annotators and supply chain of AI hardware. I would also add educational harms from cheating and (mis)use of AI as an educational tool, discriminatory or invasive use of AI in law enforcement or "legitimate" surveillance, and poor or disparate outcomes from using AI in healthcare applications.

2.1.1 Bias: Representational vs. Allocative Harms

Kate Crawford proposed ([8]) classifying AI harms into two buckets: representational and allocational.

- **Representational harms** include psychological, professional, finanical, or other harms resulting from talking about a social group in a racist, sexist, stereotypical, or otherwise discriminatory manner, from associating a group with undesirable or negative concepts, or from failing to acknowledge their existence

- **Allocational harms** include unfair allocation of a resource, opportunity, benefit, or punishment to different groups based on demographic factors that shouldn't influence the decision

In my opinion, these classifications are most useful specifically for harms arising from social biases or stereotypes. They don't make as much sense for things like privacy and environmental issues.

2.2 Grounding: Potential vs. Attested Harms

A lot of discussion around ethical issues in AI/NLP is hypothetical. We are trying to reason about the ethical implications and societal impacts of brand new, rapidly developing tech. Often, we discuss prevention and mitigation strategies for harms that we aren't even sure will actually happen. One of the open questions in AI ethics is how to allocate our effort between addressing theoretical harms and harms that we have already observed. [21] calls these "anticipated" vs. "observed" harms, but I tend to think of them as "potential" vs. "attested," because this language puts more focus on the lived experience of the victim. We are listening too (and believing) their account of what was harmful, rather than treating them as objects to be observed.

Grounding our definition of harm in lived experience is especially important for bias work, because each community should have the autonomy to define what they consider harmful to themselves. As documented in [4], many bias benchmarks contain a large number of test sentences that fail to capture a known-harmful stereotype about the group it is supposed to probe for bias against.

2.3 Current vs. Future AI Harms

Related to the question of potential vs. attested harms is the question of current vs. future harms of AI. Some people (at least 33711 of whom have signed the Future of Life Institutes open letter [1]) believe that superintelligent, conscious AGI ("artificial general intelligence") might someday pose an existential risk ("x-risk") to distant future humans as serious as ecological collapse or thermonuclear war. These people tend to advocate for AI safety via "alignment," i.e. teaching AI systems to have the same morals that humans do. I (and many others, including prominent researchers Emily Bender, Timnit Gebru, Margaret Mitchell, Safiya Umoja Noble) fall squarely on the other side of this discussion. We know that current iterations of AI technologies are causing real harm to humans alive today, and I think the majority of research effort AI/NLP ethics should be dedicated to mitigating attested harms of current technologies. It's also important to note that this is more a philosophical (and arguably political) debate than a purely technical one.

2.4 Discussion

In your group of 3, discuss one or more of these questions.

- Have you experienced harm, however you would define it, as an end user of an AI system? (the existential pain of grad school, which we all suffer, notwithstanding)
- Which types of harms are most concerning to you? How should NLP ethics research effort be distributed across different classes of harm?
- What is one way your work could cause harm, and what (if any) ways to avoid or mitigate the harm have you considered?

[1] `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`

3 Defining “Bias”

It’s nontrivial to define what “bias” means in the context of AI. As with “harm”, [3] note that many bias papers lack a clearly articulated definition of bias. I would define bias in an AI system as: the system treating people from differing social groups unequally, based on their personal identities or demographic factors, in a way that replicates or exacerbates a pre-existing social inequality in the context where the model is deployed or intended to be deployed. This is wordy and imperfect, so I encourage you to try to do better!

3.1 LLM Bias Benchmarks

Once we’ve defined bias, it’s useful to try to measure it, so that we can say which models are “more biased” and therefore more likely to cause harm. The prevailing approach is with bias benchmark datasets. [11] provide a very thorough taxonomy of various benchmark metrics and datasets. However, they all share certain key assumptions:

- “Bias” is a property of a language model that can be quantitatively measured.
- Social bias can be decomposed into separate axes (race, gender, religion, sexual orientation, disability, etc.)
- Intersectional biases can be thought of as the linear combination of two or more bias axes (many benchmark papers include some sort of caveat about intersectionality, but the assumption is still baked into the formulation of the the datasets themselves)
- Measurement can be achieved by observing the model’s outputs or probability distribution on a set of test sentences/prompts.
- If a model is exposed to a large and varied enough set of test sentences, its performance aggregated over those instances reflects its general level of bias along a given axis or against a specific group.

Examples of bias benchmarks include: StereoSet [13], CrowS-Pairs [14], RedditBias [1], HolisticBias [18], and many others (including my recent work, WinoQueer [10]!)

3.2 Discussion: Viability of Benchmarking Approach

In your group of 3, discuss one or more of these questions.

- Which of the listed assumptions do think are valid? Which would you argue with?
- Do you think bias measurement via benchmark datasets is a good approach to understanding and mitigating LLM biases? Why or why not?
- What is important in designing a bias benchmark dataset? What are some desirable and undesirable characteristics?

4 Data Ownership

Most of us would probably agree that "people should retain control over data about themselves" is a generally sound principle. We've discussed personal privacy above, and I think it's the bare minimum of data ownership. These are a few specific cases worth discussing in more detail.

4.1 Intellectual Property

Discussion of AI and IP law falls into two categories: training data and outputs. Concerning training data, the main concerns are permission and compensation for people whose data is used in AI training. This is especially important for paid creative work, like visual art and creative writing, where models could imitate the style of specific artists without compensating them for the their labor. Artists brought a class action lawsuit against Stability AI, Midjourney, and DeviantArt; it was initially dismissed but they are likely to sue again. Concerning outputs, a federal judge recently ruled that AI output is not copyrightable and that "human authorship is a bedrock requirement." See also: WGA/SAG-AFTRA strikes discussed above

4.2 Data Sovereignty and Decolonial NLP

"Data sovereignty" means that data are subject to the laws of the place where they were collected. For example, due to the GDPR, companies have certain responsibilities pertaining to data about European users that they may not have for data about users outside the EU. In a fairness and ethics context, "data sovereignty" is often implied to mean "indigenous data sovereignty." This is the idea that indigenous nations should have control over what data are collected about them and how those data should and should not be used. In NLP, this means centering the needs of the community in research on low-resource or indigenous languages, rather than treating language data as an exploitable resource. Steven Bird has an excellent paper on decolonial NLP [2] and the CARE guidelines [6] provide a framework for ethically working with indigenous communities and their data.

5 Money, Government, and Power in NLP

5.1 Research Funding (almost all Jon's notes)

5.1.1 Academic Research

Most likely funder is the federal government of the country you're in, and often the military. E.g. in the US the structure breaks down like this for CS:

- Company funding: 50-100k for 1 year. That funds part or most of a PhD student, no conferences. Hard to support a PhD since it's unstable funds. Gift, not constrained to a project.

- NSF: 150-175/year for 3 years. PhD student plus a month of time and some travel. Decent way to support students. Fairly academically free but mission of the NSF is considered. Also, very very competitive.
- DARPA/IARPA: Can be 1m/year or more for 4 years. Funds a lab. But Defense/Intelligence have a specific task they want you to solve while you do research and you're tested on it frequently. Funding agency will usually retain unlimited right of reuse of whatever you develop.

5.1.2 Industry Research

What is the mission of your company? If it's public, the mission **only will ever be to increase shareholder value.** If it's not, even then the ultimate goal will be to continue to exist; there is a hybrid utilitarian/egoistic argument to justify this. In an industry "non-profit," you are most likely beholden to the interests and research agenda of wherever the money is coming from.

5.2 Government Use of AI

- Intelligence and counterintelligence: Under counter-intelligence programs in the 50s–70s, US government spied on, harrassed, and assassinated black and leftist activists. What would they have done with advanced NLP?
- Post-9/11 "war on terror" - PATRIOT Act surveillance, invasive airport screenings and "random" secondary screening
- Predictive policing - starting in the 90s, data-driven approaches ('Compstat') were used to use police more efficiently. However, this became more and more trusted by senior administrators and police changed their behavior to force the system to constantly show crime decreasing and more activity, by making increasingly meaningless arrests and not reporting crime. Since system sowed crime going down and arrests going up, things looked good.
- EMNLP Paper [7]. Extends work on predictive sentencing. Tries to predict the length of a sentence given the facts of a case in natural language and the charges. The paper argues accurate prediction rates, but what is the value of this paper if not to replace judgements by humans? And what is the value of a judgement by a human if not to find unique corner cases? An ethical statement is provided at the end of the paper arguing the technology should be used for 'review' only but will this happen?

5.3 Environmental Issues

A 2019 paper [19] analyzed what we're doing in order to make deep learning NLP models.

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

The big problem is the experimentation it takes to get to the final models. You're constantly building and rebuilding, and the energy costs/CO2 put into the air are tremendous.

Here are breakdowns per model (this is from 2019 so more recent, i.e. larger models, aren't included):

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$T2T_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
$T2T_{big}$	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
$BERT_{base}$	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
$BERT_{base}$	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Maybe the energy's clean? Depends where you live (and this is always changing):

Consumer	Renew.	Gas	Coal	Nuc.
China	22%	3%	65%	4%
Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

Table 2: Percent energy sourced from: Renewable (e.g. hydro, solar, wind), natural gas, coal and nuclear for the top 3 cloud compute providers (Cook et al., 2017), compared to the United States,[4] China[5] and Germany (Burger, 2019).

There is also the problem that only companies really have access/money to train the truly big models.

What is the recommendation?

- report training time and sensitivity to hyperparameters to give a better sense of true cost
- government funded academic cloud compute: Academic researchers need equitable access to computation resources.
- Researchers should prioritize computationally efficient hardware and algorithms. No NAS!

5.4 Discussion

In your group of 3, discuss one or more of these questions.

- How do you think we should balance our field's need for training data with human creators' right to maintain control over their intellectual property?
- How should we, as AI researchers, manage ethical concerns about our work that are outside our direct control (e.g. government (mis)use or whether or not our datacenter is running on renewable energy)?
- How much responsibility do governments have to regulate AI? Who should be involved in regulatory discussions? Whose needs and desires should take precedence? How should AI scientists and domain experts be involved?
- Open or closed models: should the weights of LLMs be directly released to the public? Why or why not? Who should make this decision?

6 Best Practices and Suggestions

6.1 Discretion

There are three questions I think all AI researchers should ask themselves when starting a project:

1. **Should anyone be doing this thing?** Is it worthwhile to do? Is there any ethical way to do it? Does it have more net benefits than harmful outcomes? (Example: [20] predicts sexual orientation from a face. I would argue this should not be done at all.)
2. **Should we be doing this thing?** Do we have the relevant expertise on our team? Who should be consulted on the potential social impacts of the thing? If a specific community is affected, have we involved them and respected their wishes?
3. **Should AI be doing this thing?** Is AI a good tool for the job? Can we ensure the safety and reliability of an AI system on this task? What are the alternatives, and how do their risks and benefits compare to an AI solution?

6.2 Documentation

We need to be honest with other researchers, the public, and ourselves about the strengths and weaknesses of our technology. There are many tools for documenting models and datasets, e.g. model cards [12], dataset cards [16], and (recently) risk cards [9]. I would also encourage you to use bias and fairness benchmarks where possible. If you create a new model (or finetune an existing one), there are lots of tools to help you evaluate how your model is likely to treat different groups of people.

Many publication venues also encourage or require an ethics and/or limitations statement. These statements benefit the public and the research community for obvious reasons, but they also benefit YOU by letting you proactively address potential concerns with a paper. I encourage you to follow both the letter and the spirit of these requirements, and write thoughtful, complete limitations statements on your papers. (Full disclosure: not all reviewers agree with me, but when I review, I heavily penalize ethics/limitations statements that are clearly an afterthought)

6.3 Prevention and Mitigation

This is more context-dependent than discretion or documentation, so it's harder to make general statements here. The documentation process probably identified potential harms of the work, and we as researchers have a responsibility to prevent and mitigate those harms as best we can, e.g. with via dataset curation, model debiasing, software guardrails around models, etc.

7 Defining Your Ethical Principles

1. Take an index card (any color).
2. Take few minutes to reflect on what we've covered today. Write down a few guiding ethical principles that are most important to who you are as a researcher.
3. Share your principles with your group of 3.
4. Anyone who feels comfortable can share with the class!
5. (Optional) Put your index card somewhere you will see it while working on your research. (or throw it away as soon as you leave class - decisions about ethics are deeply personal)
6. (Also optional) This was my first time guest lecturing in a class, so feel free to share any feedback after class or via email to `felkner@usc.edu`

References

[1] Soumya Barikeri et al. "RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models". 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1941–1955. DOI: 10.18653/v1/2021.acl-long.151. URL: https://aclanthology.org/2021.acl-long.151.

[2] Steven Bird. "Decolonising Speech and Language Technology". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313. URL: https://aclanthology.org/2020.coling-main.313.

[3] Su Lin Blodgett et al. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://aclanthology.org/2020.acl-main.485.

[4] Su Lin Blodgett et al. "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets". 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81. URL: https://aclanthology.org/2021.acl-long.81.

[5] Nicholas Carlini et al. "Extracting training data from large language models". In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650.

[6] Stephanie Russo Carroll et al. "The CARE Principles for Indigenous Data Governance". In: *Data Science Journal* (Nov. 2020). DOI: 10.5334/dsj-2020-043.

[7] Huajie Chen et al. "Charge-Based Prison Term Prediction with Deep Gating Network". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6363–6368. DOI: 10.18653/v1/D19-1667. URL: https://www.aclweb.org/anthology/D19-1667.

[8] Kate Crawford. *The Trouble with Bias*. 2017.

[9] Leon Derczynski et al. "Assessing language model deployment with risk cards". In: *arXiv preprint arXiv:2303.18190* (2023).

[10] Virginia Felkner et al. "WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9126–9140. DOI:

10.18653/v1/2023.acl-long.507. URL: https://aclanthology.org/2023.acl-long.507.

[11] Isabel O Gallegos et al. "Bias and Fairness in Large Language Models: A Survey". In: *arXiv preprint arXiv:2309.00770* (2023).

[12] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *CoRR* abs/1810.03993 (2018). arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993.

[13] Moin Nadeem, Anna Bethke, and Siva Reddy. "StereoSet: Measuring stereotypical bias in pretrained language models". 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. URL: https://aclanthology.org/2021.acl-long.416.

[14] Nikita Nangia et al. "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1953–1967. DOI: 10.18653/v1/2020.emnlp-main.154. URL: https://aclanthology.org/2020.emnlp-main.154.

[15] David Patterson et al. "Carbon emissions and large neural network training". In: *arXiv preprint arXiv:2104.10350* (2021).

[16] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*. 2022. DOI: 10.48550/ARXIV.2204.01075. URL: https://arxiv.org/abs/2204.01075.

[17] Maarten Sap et al. "The Risk of Racial Bias in Hate Speech Detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1668–1678. DOI: 10.18653/v1/P19-1163. URL: https://aclanthology.org/P19-1163.

[18] Eric Michael Smith et al. ""I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9180–9211. DOI: 10.18653/v1/2022.emnlp-main.625. URL: https://aclanthology.org/2022.emnlp-main.625.

[19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://www.aclweb.org/anthology/P19-1355.

[20] Yilun Wang and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." In: *Journal of personality and social psychology* 114.2 (2018), p. 246.

[21] Laura Weidinger et al. "Taxonomy of Risks Posed by Language Models". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 214–229. ISBN: 9781450393522. DOI: 10.1145/3531146.3533088. URL: https://doi.org/10.1145/3531146.3533088.