Ethics and Power in NLP

Jonathan May

September 28, 2022(Prepared for Fall 2022)

There are entire courses on ethics in AI and NLP. A list of them is here: `https://aclweb.org/aclwiki/Ethics_in_NLP`. So this lecture will necessarily be incomplete. I'm also highly influenced here by a talk Alvin Grissom II (Ursinus College) gave at WiNLP in Summer, 2019. Also see Fairness in ML tutorial `https://mrtz.org/nips17/#/`. And Tsvetkov/Black course `http://demo.clab.cs.cmu.edu/ethical_nlp/`. I've been also influenced by a lecture Batya Friedman gave on Value Sensitive Design (`https://www.envisioningcards.com/`).

However, **The Views In These Lecture Notes are Entirely My Own**

1 Ethics

What does it mean to you?

- Societal Standards of right and wrong (though you can have ethics that you think are correct but that go against the grain of the current society, which you would then consider to be ethically corrupt. See, e.g. nazi germany)
  - Things that should be promoted (honesty? compassion? loyalty? creative production?)
  - Things that should be discouraged (murder? theft? injury?)
- Consideration of how important forces deviate from these standards.
  - personal feelings (i want to kill that person)
  - laws
  - social norms

Here are some ways of breaking approaches to ethics down

- Deontological approaches – Consider certain actions themselves to be simply good or bad. There are a set of universal rules and you follow them.
  - Moral absolutism – universal moral principles that exist. When these get rigidly followed there can be consequences nobody (or most people) likes

  - See also Asimov's laws of robotics: 1. Don't harm humans, 2. Follow humans. 3. Don't harm yourself. But then how do robot police deal with serial killers?

- Consequentialist approaches – judge whether something is right based on what the outcome is
  - Issue with this is sometimes you don't know what that will be.
  - Utilitarianism – do whatever provides greatest good for greatest number of people (where 'good' = knowledge/pleasure/health/aesthetics). This takes society into account but can lead to some pretty awful behaviors. It also assumes that people making the decisions know what is best for others.
  - Egoism – everybody works in their own self-interest. Assumes everyone has the capacity to equally work in their self interest. Also doesn't really have a solution for

So the upshot is this is not easy. I would advocate for a mix. Guidelines (so mildly deontological) but with rational approaches and corrections based on actual outcomes (so mildly consequentialist) with an eye toward helping most (so somewhat utilitarianistic) but encouraging this to be conveyed in a bottom up way (so somewhat egoistic).

2 Some Principles We May Want to Support

I want to remind the reader again that these notes reflect my (Jon's) opinion and don't reflect any institution I work for or don't work for! The following strike me as a reasonable set of guidelines to strive for:

- **Groups** as noted below include but are not necessarily limited to gender identity, gender expression, sexual orientation/identity/non-identity, disability, marital/family status, race/ethnicity, class, language use, politics, religion, age, national origin
- Groups should be able to choose to use a technology to its fullest potential, in a way that benefits them.
- Groups should be able to reject the use of a technology without negative consequences apart from not receiving direct benefit from the technology.
- Groups should know if they are being affected by a technology even if they aren't making direct use of it.
- Groups should have autonomy over their personal information and the ability to keep that information private.
- Groups should be fully informed of the trade-offs of the use of their information by a technology and the benefit others may receive.
- Groups should understand what a technology can and what its limits are.

Some examples to consider/discuss:

- use to fullest potential: translation not speaking your language, speech recognition not recognizing your accent, your dialect not handled by too-rigid NLU
- rejection without consequences: not getting a job if you're off of twitter or linkedIn.
- affected without using a tech: Predictive technology making heteronormative biases in sentence completion (e.g. 'I'm such a lucky girl I just got engaged to my ...') can be othering, can reinforce biases. See also section 4.
- autonomy over personal information/privacy: Scraped tweet corpora that are redistributed as plain text; Clearview AI scraped internet for face images, built massive recognition DB, sold tech to law enforcement
- Informed of trade-offs of use of info: social media companies can sell the data you allow them to collect to other companies for their marketing purposes (unless you opt out, now)
- understand limits: can meaning be learned from only form ? [1]

Exercise: Do one of the following:

1. Consider the application topic of your reproduction study. What are unintended consequences that could arise out of the application if it is successful? Or, what are populations that might not benefit equally from the application, and why?
2. Consider a technology that displays a wearer's emotion on a t-shirt they wear. What are some consequences that could come of children using the technology?

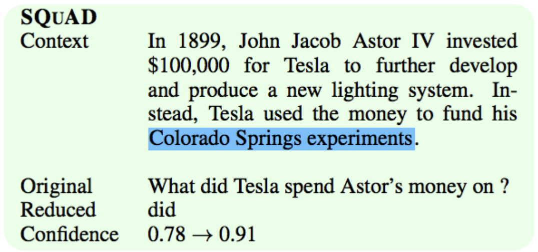
3 Problems in how classification is achieved

It seems that we shouldn't really care *how* a classifier works, we should only care *that* it works, and if a classifier works better, it's a better classifier.

A counter argument is that classification, aka discrimination, is appropriate when it is domain specific, not general. When irrelevant or, more importantly, historically unjustified/systematically adverse results have been used for classifying, we can say (deontologically) that we should stop using the current approaches.

3.1 Uncharged example: question answering with the wrong signal

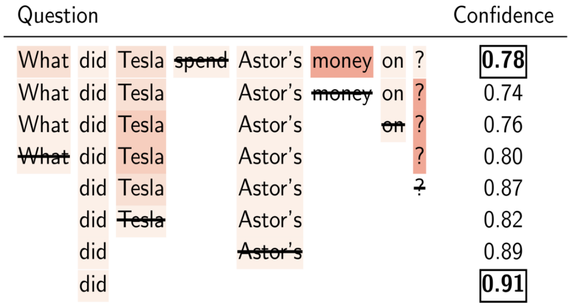
The work referenced is [5].



SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	0.78 → 0.91

Figure 1: SQuAD example from the validation set. Given the original *Context*, the model makes the same correct prediction ("Colorado Springs experiments") on the *Reduced* question as the *Original*, with even higher confidence. For humans, the reduced question, "did", is nonsensical.



Question	Confidence
What did Tesla ~~spend~~ Astor's money on ?	$\boxed{0.78}$
What did Tesla Astor's ~~money~~ on ?	0.74
What did Tesla Astor's ~~on~~ ?	0.76
~~What~~ did Tesla Astor's ?	0.80
did Tesla Astor's ~~?~~	0.87
did ~~Tesla~~ Astor's	0.82
did ~~Astor's~~	0.89
did	$\boxed{0.91}$

Hopefully you agree that the wrong info is being used to make the right choice. And furthermore that this could very well lead to the wrong info being used to make the wrong choice.

3.2 More charged: Race and gender bias in NLP

[3]: pretty much every kind of bias you can imagine was observed in glove embeddings. Typical European-American names associated with pleasant words; black American names associated with negative words. Typical names for woman associated with arts; those for men associated with science.

Why is this a problem? For one thing, having stereotype biases, particularly strongly weighted ones, in your models, can lead to your models predicting the wrong thing, even if evidence beyond the bias counters the biased output.

Example: winograd test with bias potential [9]:

Consider these sentences:

1. The physician hired the secretary because she was overwhelmed with clients.
2. The physician hired the secretary because he was overwhelmed with clients.
3. The physician hired the secretary because she was highly recommended.
4. The physician hired the secretary because he was highly recommended.

Does your model prefer 2 over 1 and 3 over 4? Moreover if you have a sentence fragment 'The physician hired the secretary because she' can your model resolve the pronoun with high confidence? That's indication of bias based on priors that's not paying attention to the language of the sentence itself (a kind of *posterior collapse*).

How do we:

- ...determine there is this problem in the first place?
- ...solve the problem?

To detect the problem it's hard with modern models to isolate the signals being used. We can try with attention (see in QA example above) but that's an imperfect window as well. More typical is to create 'debiased' test sets which specifically probe for 'nontraditional' outputs. For example, in [9], default models evaluating on the 'cross-bias' set are on average 21.1 worse in F1.

How about to solve the problem? I think a number of groups at USC are working on this so I'd like to hear from the experts here. But I know that data augmentation (swap stereotypical entities in training data) mitigates somewhat...but often only in one dimension at a time. Gender is not binary, though binary gender dominates data and discussion. And what about e.g. race – much more than binary and more balance in this regard. But assumptions like binary gender and black vs. white are common and miss lots of nuance in the way bias exists.

A counter argument to trying to address the problem is that it isn't a problem. E.g. 'people are biased, we're just reflecting the data.' It's helpful to consider where the biases come from in the first place. The language choice perpetuates stereotypes: Consider an article about a black man stabbed by a white supremacist and how it ran in the New York Post:

> Caughman, **who has 11 prior arrests**, walked for about a block after the stabbing and staggered into the Midtown South Precinct, looking for help. He died hours later after being rushed to a nearby hospital. Police sources said the **career criminal was refusing to talk to police about the incident and acting combative before his death.**

This kind of loaded language (in the added emphasis) is not atypical from the NYPost, and likely comes out of harbored biases that are propagated by refusing to combat what is printed in places like the NYPost. What is read in news articles is not typical of *anybody's* lived experience because if it was *it wouldn't be news!!*

Incidentally, this bias is not necessarily limited to 'known offenders (e.g. NYPost).' [2] argues that you can't really create something without some intentionality:

> A former Apple employee...described his experience on a team that was developing speech recognition for Siri... As they worked on several English dialects, he asked his boss: "What about African American English?" To which his boss responded: "Well, Apple products are for the premium market."

3.3 Unintentional effects

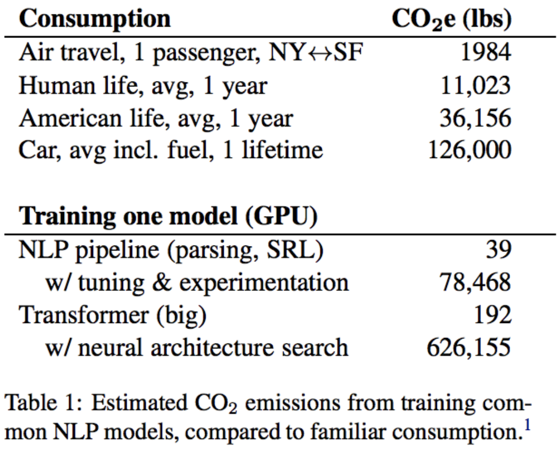
COMPAS – a system for predicting probability of criminal reoffending. It was trained on a balanced data set, and race was not an input feature. However, ZIP code was, ZIP is highly correlated to race in the US, because of historical housing discrimination policies. Race is also highly correlated to socioeconomic difficulty, for the same reasons.

Additionally, the data was set up to predict whether a person would **commit a serious crime**. How was this judged? By who is likely to be **convicted**. Conviction rates are also correlated strongly with race.

We can talk about algorithms to debias these results. But people have to want to use them. If you're trying to get a new SOTA on a GLUE task, and being biased helps because the *test set is biased*, what is the right move?

4 Power, i.e. Energy

A recent paper [8] analyzed what we're doing in order to make deep learning nlp models.

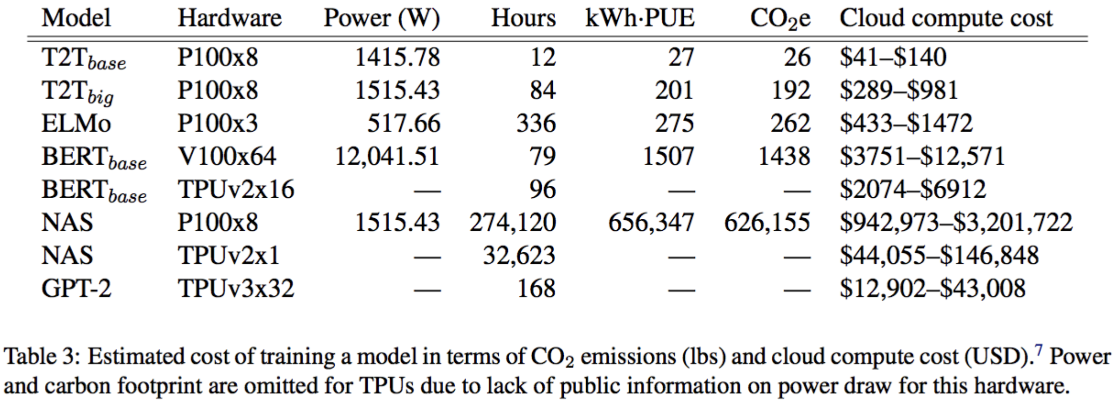


Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture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

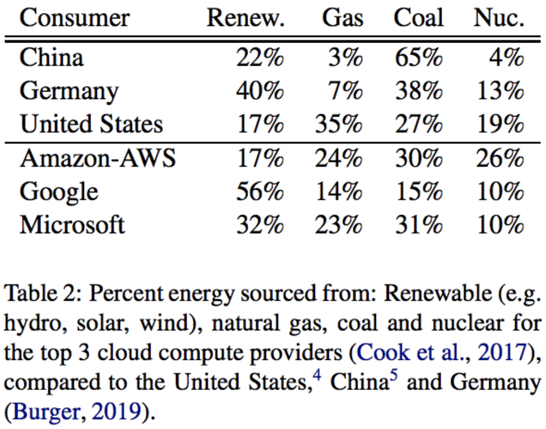
Here are breakdowns per model:



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:



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. 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 Power, i.e. Control

5.1 Who funds your research?

5.1.1 In a University?

Then probably 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.

Unlimited rights of reuse are generally given to the funding agencies (esp. DARPA/IARPA). So be careful what you develop!

- Under counter-intelligence programs in the 50s–70s, US government spied on, harrassed, and assassinated black and leftist activists
- FBI currently targeting ”black identity extremists”
- What would they do with advanced NLP?
- Consider treatment of MLK by FBI under Hoover

5.1.2 In a company?

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.

It's hard to avoid being results-driven and the evidence shows that's what continues to happen:

- face recognition false positives on white male faces way less than other combinations. Do we expect this to be any different if detecting social media text and predicting malfeasance?

5.2 How will your research be used to exert power over others?

- 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 [4]. 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 Codes of Ethics

From Hal Daume (2016).

5.3.1 IEEE:

1. to accept responsibility in making decisions consistent with the safety, health, and welfare of the public, and to disclose promptly factors that might endanger the public or the environment;
2. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
3. to be honest and realistic in stating claims or estimates based on available data;
4. to reject bribery in all its forms;
5. to improve the understanding of technology; its appropriate application, and potential consequences;
6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
7. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others;
8. to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression;
9. to avoid injuring others, their property, reputation, or employment by false or malicious action;
10. to assist colleagues and co-workers in their professional development and to support them in following this code of ethics.

5.3.2 From Hal:

Responsibility to the Public:

1. Make research available to general public
2. Be honest and realistic in stating claims; ensure empirical bases and limitations are communicated appropriately
3. Only accept work and make statements on topics which you believe have competence to do
4. Contribute to society and human well-being, and minimize negative consequences of computing systems
5. Make reasonable effort to prevent misinterpretation of results

6. Make decisions consistent with safety, health and welfare of public
7. Improve understanding of technology, its application and its potential consequences (positive and negative)

Responsibility in Research:

1. Protect the personal identification of research subjects, and abide by informed consent
2. Conduct research honestly, avoiding plagiarism and fabrication of results
3. Cite prior work as appropriate
4. Preserve original data and documentation, and make available
5. Follow through on promises made in grant proposals and acknowledge support of sponsors
6. Avoid real or perceived COIs, disclose when they exist; reject bribery
7. Honor property rights, including copyrights and patents
8. Seek, accept and offer honest criticism of technical work; correct errors; provide appropriate professional review

Responsibility to Students, Colleagues, and other Researchers:

1. Recognize and property attribute contributions of students; promote student contributions to research
2. No discrimination based on gender identity, gender expression, disability, marital status, race/ethnicity, class, politics, religion, national origin, sexual orientation, age, etc.
3. Teach students ethical responsibilities
4. Avoid injuring others, their property, reputation or employment by false or malicious action
5. Respect the privacy of others and honor confidentiality
6. Honor contracts, agreements and assigned responsibilities

Compliance with the code:

1. Uphold and promote the principles of this code
2. Treat violations of this code as inconsistent with membership in this organization

5.3.3 Deontological elements specific for NLP/linguistics

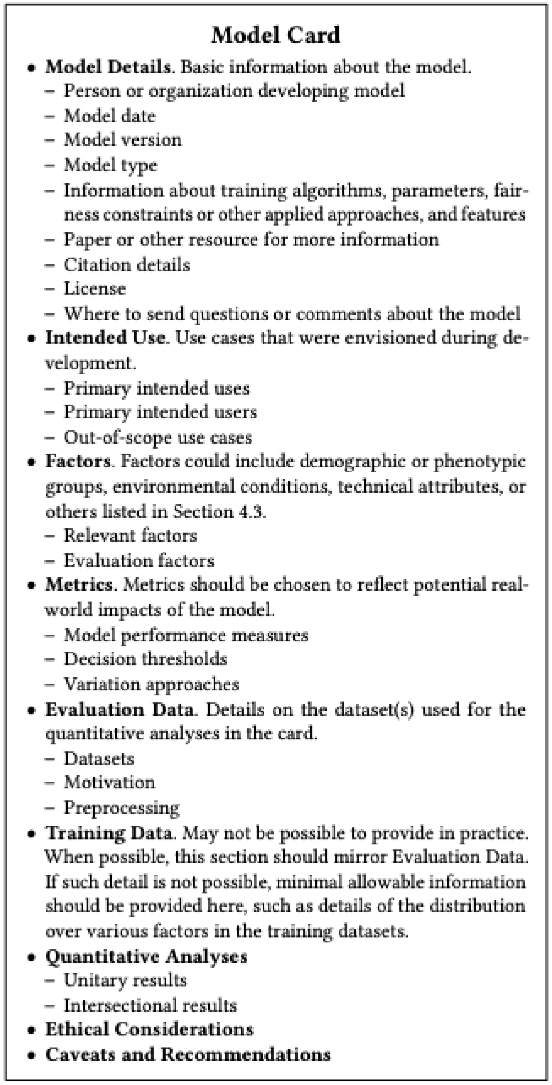
Support language variability and diversity

Recognize and model language as it is used

Respect the rights of humans to keep private language private

6 Responsibility in Reporting

6.1 Model Cards[6]



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

The idea is to show what a model was designed to do and what it wasn't designed to do. Specifics about when the model was built, what version of a codebase it represents, and what referring papers to consult are specified. Data the model was trained on and how it was evaluated are included. Various risks should be noted as well as ethical considerations.

Apart from helping users of the model this seems worthwhile for the people that write the model card.

They're pretty heavily used on huggingface: https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (kind of hard to display as image in notes).

6.2 Data Cards[7]

Like a model card but for the creation and life cycle of data sets. I am in general more interested in the challenges here but actually this paper is a lot less mature than model cards so for now there's not too much info here.

6.3 Statements

Increasingly there are ethics statements, limitations statements, ethics checklists etc. mandated at the end of papers or submitted with papers. It's unclear so far if they're having the full disclosure effect or are changing how we do research, but they do encourage thought.

References

[1] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

[2] Ruha Benjamin. *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons, 2019.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[4] 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.

[5] Shi Feng et al. "Pathologies of Neural Models Make Interpretations Difficult". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018). DOI: 10.18653/v1/d18-1407. URL: http://dx.doi.org/10.18653/v1/d18-1407.

[6] 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.

[7] 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.

[8] 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.

[9] Jieyu Zhao et al. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (2018). DOI: 10.18653/v1/n18-2003. URL: http://dx.doi.org/10.18653/v1/N18-2003.