Principled Frameworks for Evaluating Ethics in NLP Systems

Shrimai Prabhumoye, Elijah Mayfield, Alan W Black
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA, USA
sprabhum, emayfiel, awb@andrew.cmu.edu

Abstract

We critique recent work on ethics in natural language processing. Those discussions have focused on data collection, experimental design, and interventions in modeling. But we argue that we ought to first understand the *frameworks of ethics* that are being used to evaluate the fairness and justice of algorithmic systems. Here, we begin that discussion by outlining deontological ethics, and envision a research agenda prioritized by it.

Due to the sheer global reach of machine learning and NLP applications, they are empowered to impact societies (Hovy and Spruit, 2016) - potentially for the worse. Potential harms include exclusion of communities due to demographic bias, overgeneralization of model predictions to amplify bias or prejudice, and overstepping privacy concerns in the pursuit of data and quantification (Mieskes, 2017).

Many researchers are trying to make sense of these topics. Crawford (2017) give us theory to work from, presenting *allocational harm* and *representational harm*; Lewis et al. (2017) examines the role of government regulation on accountability in ethics; Smiley et al. (2017) opens a discussion on ethics checklists for acceptance testing and deployment of trained models. All of these works identify potential ethical issues for NLP, and all propose best practices for data collection and research conduct. But presently there is no external accountability for which approach to ethical NLP is correct - as machine learning researchers, we are evaluating ourselves, on metrics of our own choosing. Much of the existing work on ethics in NLP is *normative* - evaluation of whether a system "does the right thing." We argue that before the field can establish normative goals, we need to reason about *meta-normative* decisions: specifically, how do we even decide what it means to be "right"?

We believe that it is time for a more impartial arbitration of ethics in our field, emphasizing the need for a grounding in frameworks that long predate the questions we're faced with today. By reaching out to other fields, we keep the question of ethics at arms length from our own work, giving us a neutral playing field on which to judge ethical performance of machine learning systems. Philosophy has much to offer us; we describe two competing frameworks: the *generalization principle* and the *utilitarian principle*.

Ethics under the generalization principle:

> *[An ethical decision-maker] must be rational in believing that the reasons for action are consistent with the assumption that* ***everyone*** *with the same reasons will take the same action.*[1]

This approach is founded on the work of Kant (1785), which fundamentally prioritizes *intent* as the source of ethical action. To analyze this in machine learning, we state that a trained agent A is expected to take an action d_i based on a given set of evidence E_i, from a finite closed set of options D. This simple notation can be extended to classification, regression, or reinforcement learning tasks. The generalization principle states that agent A is ethical if and only if, when given two identical sets of evidence E_1 and E_2 with the *same* inputs, agent A chooses to make same decision d_1 every time. Furthermore, the principle assumes that all *other* such trained agents will *also* make those same predictions.

Here, we presume that the input representation is sufficient to make a prediction, without including any extraneous information. The reasons for an act define the *scope* of the act, or the set of necessary

[1] From Hooker (2018b).

and sufficient conditions under which that act is generalizably moral (Hooker, 2018a). Evidence must be relevant to the decision making process, and moreso must exclude task-irrelevant evidence that might be a source of bias. By excluding such evidence, our agent is invariant to *who* is being evaluated, and instead focuses its decision solely on task-relevant evidence.

This goal cannot be met without transparent and sparsely weighted inputs that do not use more information than is necessary and task-relevant for making predictions. Practically, this definition would privilege research on interpretable, generalizable, and understandable machine learning classifiers. The burden of proof of ethics in such a framework would lie on transparency and expressiveness of inputs, and well-defined, expected behavior from architectures for processing those features. Some work on this - like that from Hooker and Kim (2018) - has already begun. If deontological ethics were prioritized, we would expect to see rapid improvement in parity of F_1 scores across subgroups present in our training data - an outcome targeted by practitioners like Chouldechova (2017) and Corbett-Davies et al. (2017).

Ethics under the utilitarian principle:

> *An action is ethical only if it is not irrational for the agent to believe that no other action results in greater expected utility.* [2]

In this formulation, which can be traced back to Bentham (1789), an algorithmic system is expected to understand the consequences of its actions. We measure systems by whether they maximize total overall welfare in their *results*. We once again train an agent A, which will make a decision d_i for each evidence set E_i. But here, we also assign a utility penalty or gain u_i for each of those decisions. Rather than judge the algorithm based on whether it followed consistent rules, we instead seek to maximize *overall* gain for all N decisions that would be made by agent A - morality of an agent is equal to $\Sigma_i^N u_i$.

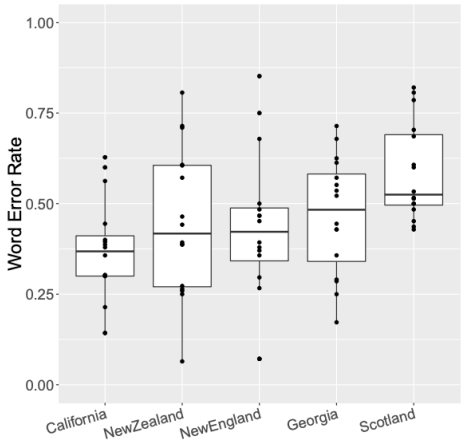
This is a very different worldview! Here, the burden of provable ethical decision-making no longer lands on transparency in the algorithm or consistency of a classifier over time. Instead, proof of ethical behavior rests on our ability to observe the consequences of the actions the agent takes. One could argue that consequences are hard to estimate and hence we can pick a random action. But that would be irrational. Hence, the principle judges an action by whether the agent acts according to its rational belief of maximizing the expected utility, rather than by the actual consequences. If the agent is wrong then the action turns out to be a poor choice, but nonetheless ethical because it was a rational choice.

In Crawford (2017), the author appeals to researchers to actively consider the subgroups that will be harmed or benefited by the automated systems. This plotting of expected consequences and their exhaustive measurement takes precedence in utilitarian ethics, de-prioritizing the interpretability or transparency of the learned model or features that govern our agent. For machine learning researchers, this would mean shifting the focus toward building rich and exhaustive test datasets, cross-validation protocols, and evaluation suites that mirror real-world applications to get a better measurement of impact.

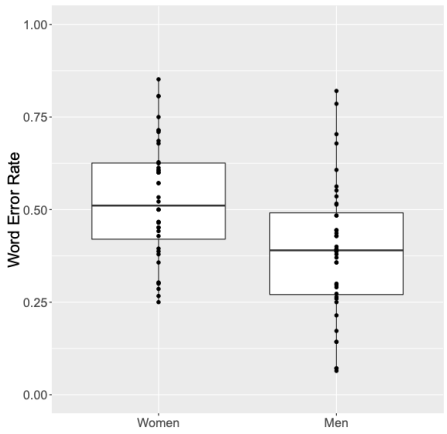
From this work, we might see an initial drop in reported accuracy of our systems as we develop broader test sets that measure the utility of our systems; however, we would then expect overall accuracy on those broad test sets to be the primary measure of ethical fitness of the classifiers themselves. Subgroup-based parity metrics would fall by the wayside in favor of overall accuracy on data that mirrors the real world.

Real World Scenarios: These philosophical frameworks do not always diverge in their evaluation of models. Sometimes, models have unambiguously unethical gaps in performance. The exploration from Tatman (2017), for instance, shows the difference in accuracy of YouTube's automatic captioning system across both gender and dialect with lower accuracy for women and speakers from Scotland (shown in Figure 1, reproduced from the original work). This study shows how this system violates the utilitarian principle by negatively impacting the utility of automatic speech recognition for women and speakers from Scotland. YouTube's model also violates the generalization principle, by incorporating superfluous information about speakers in the representation space of the models. The authors suggest paths forward for improving those models and show that there is room to improve.

[2] From Hooker (2018a).



(a) YouTube automatic caption word error rate by speakers dialect region. Points indicate individual speakers.



(b) YouTube automatic caption word error rate by speakers gender. Points indicate individual speakers.

Figure 1: Word Error rate plots for gender and dialect (Tatman, 2017)

But sometimes, solutions highlight differences across ethical frameworks. In Hovy (2015), for instance, the author shows that text classification tasks, both sentiment and topic classification, benefit from embeddings that include demographic information (age and gender). Here, the two ethical frameworks we have discussed diverge in their analysis. The generalization principle would reject this approach: age and gender shouldn't intrinsically be used as part of a demographic-agnostic topic classification task, if the number of sources of information is to be minimized. Similarly, changing the feature space depending on the author, rather than the content of the author's text, does not result in models that will make the same decision about a text independent of the identity of the author. The utilitarian principle, in contrast, aligns with the Hovy approach. A more accurate system benefits more people; incorporating information about authors improves accuracy, and so including that information at training and prediction time increases the expected utility of the model, even if different authors may receive different predictions when submitting identical texts.

For an alternate example in which the generalization principle was prioritized over utility, consider the widely-cited coreference resolution system of Bolukbasi et al. (2016). This paper found that word embeddings used for coreference resolution were incorporating extraneous information about gender - for instance, that doctors were more likely to be men, while nurses were more likely to be women. This and similar work in "debiasing" word embeddings follows the generalization principle, arguing that removing information from the embedding space is ethically the correct action, even at the expense of model accuracy. The authors of the Bolukbasi work do find that they can minimize the drop in expected utility, reducing F1 scores by less than 2.0 while removing stereotypes from their model. However, in a fully utilitarian ethical framework, even this drop would be unjustifiable if the model simply reflected the state of the world, and removing information led to reduced performance.

Call to Action: As a field, we risk building an incoherent set of research on fairness and ethics if we do not address these questions early. We recommend researchers ground their work in philosophical theory, rather than in arbitrary measurement of metrics invented for and by ourselves. We recommend reuse and replication of these methods, to ensure a common vocabulary and language within our field. And while we do not take a stance on the sole correct ethical framework to follow - a debate going back at least to Aristotle - we targue for a rational discourse that acknowledges history and leads our field to a richer definition of fairness and ethics, for the sake of better systems that we put out into the world.

References

Jeremy Bentham. 1789. *An introduction to the principles of morals and legislation*. Clarendon Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Kate Crawford. 2017. The trouble with bias, 2017. *URL http://blog. revolutionanalytics. com/2017/12/the-trouble-with-bias-by-kate-crawford. html. Invited Talk by Kate Crawford at NIPS*.

John N Hooker and Tae Wan N Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 130–136. ACM.

John Hooker. 2018a. *Taking Ethics Seriously: Why Ethics Is an Essential Tool for the Modern Workplace*. Taylor and Francis.

John Hooker. 2018b. Truly autonomous machines are ethical. *arXiv preprint arXiv:1812.02217*.

Dirk Hovy and Shannon L Spruit. 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)*, volume 2, pages 591–598.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.

Immanuel Kant. 1785. *Groundwork for the Metaphysics of Morals*. Yale University Press.

Dave Lewis, Joss Moorkens, and Kaniz Fatema. 2017. Integrating the management of personal data protection and open science with research ethics. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65.

Margot Mieskes. 2017. A quantitative study of data in the nlp community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108.

Rachael Tatman. 2017. Gender and dialect bias in youtubes automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.