AI and Ethics (2023) 3:1185–1192
https://doi.org/10.1007/s43681-022-00226-9

ORIGINAL RESEARCH

Rawlsian AI fairness loopholes

Anna Katrine Jørgensen[1] · **Anders Søgaard**[1]

Received: 11 August 2022 / Accepted: 3 October 2022 / Published online: 26 October 2022

Abstract
Researchers and industry developers in artificial intelligence (AI) and natural language processing (NLP) have uniformly adopted a Rawlsian definition of fairness. On this definition, a technology is fair if performance is maximized for the least advantaged. We argue this definition has considerable loopholes, which can be used to legitimize common practices in AI/NLP research that actively contributes to social and economic inequalities. Such practices include what we shall refer to as Subgroup Test Ballooning and Snapshot-Representative Evaluation. Subgroup Test Ballooning refers to the practice of initially tailoring a technology to a specific target group of technology-ready early adopters to collect feedback faster. Snapshot-Representative Evaluation refers to the practice of evaluating a technology on a representative sample of current end users. Both strategies may contribute to social and economic inequalities but are commonly justified using arguments familiar from political economics and grounded in Rawlsian fairness. We discuss an egalitarian alternative to Rawlsian fairness, as well as, more generally, the roadblocks on the path toward globally and socially fair AI/NLP research and development.

Keywords Artificial intelligence · Natural language processing · Equality · Fairness · John Rawls · Kai Nielsen

1 Introduction

We begin with a thought experiment, designed to set the stage for our discussion of the global and social fairness of research and development in artificial intelligence (AI) and natural language processing (NLP). We plunge right in:

Thought experiment: the egalitarian martian Imagine a Martian visiting Planet Earth to evaluate the social impact of AI/NLP. The Martian is not interested in the *relative* social impact compared to other technologies used on Planet Earth, since the Martians would implement AI/NLP on Mars, where other technologies are used than those relied upon on Planet Earth. Imagine also that Mars—with a population of a billion Martians, roughly—is similar to Planet Earth in exhibiting linguistic diversity, with major and minor languages, but differs from Planet Earth in exhibiting perfect equality of opportunities, including income equality. The Martian has been asked by her president – Supreme Leader Xaroline – to evaluate whether AI/NLP is compatible with such equality of opportunities. Now, what would the Martian likely find?

AI/NLP refers to a vast range of technologies. We will use speech recognition as our running example: Speech recognition models today tend to be neural networks whose weights have been adjusted on millions of examples, to learn a mapping from audio of someone speaking to a text transcription of what was said. On Planet Earth, speech recognition, like many other technologies, generally works better for some languages (English) rather than others, and for some subgroups (young men) rather than others. Upon observing such a bias the Martian would likely try to identify its source. One underlying dynamic should be familiar to most observers of the field: Industry players—and to some extent, research labs—target the most ready adoption-ready groups in society–often young, urban men in the US and Europe–trying to get products out as fast as possible, leveraging the fact that resources are widely available for English, and that return on investment will presumably be larger for these groups. Only secondarily, technologies are transferred to or scaled up to other groups and languages, often in trimmed-down versions and with lower performance. In practice, transfer is slow and often gets stuck along the way due to the smaller

✉ Anders Søgaard
soegaard@di.ku.dk

[1] University of Copenhagen: Kobenhavns Universitet, Copenhagen, Denmark

revenue in smaller markets,[1] as well as companies' incentive to develop new technologies (for English) rather than to internationalize existing technologies. This, over time, leads to larger and larger performance gaps between English and other languages as well as technological scarcity for non-English languages.

Example: danish speech recognition Danish speech recognition has lacked behind for decades, and attempts to roll out speech recognition in industry or in the public sector have generally disappointed those involved. The best publicly available speech recognition model for Danish at the time of writing was developed by a multinational technology company prior to release of one of their products for the Danish market. Since the product's target group was young, urban users, they collected speech data from users of age 20–30 from Denmark's largest cities. The net result is a speech recognition model that works well if you are young and urban–and terribly, if you are not.[2]

This practice we refer to as **Subgroup Test Ballooning**, i.e., the practice of initially tailoring a technology to a specific target group of technology-ready early adopters to collect feedback faster.[3] When researchers and representatives of industry defend this practice, they typically resort to the following narrative: We develop speech technologies on English and for young, urban end users, because we have the English resources to test technologies with limited costs, enabling us to explore a wider range of technologies, to the eventual advantage of *all* potential end users, and because young end users provide fast turn-around through frequent and efficient feedback. Fast turn-around means rapid development, again to the advantage of *all* end users.

The problem with this narrative is, of course, that there is little evidence to support that low cost exploration and fast turn-around benefit out-of-target end users. If that were the case, transferring an existing technology to a new audience would be plug-and-play as soon as the data had been collected. As most practitioners will know, this is not the case. Market differences, linguistic differences, as well as differences between the needs and preferences of different groups of end users, complicate this transfer of technologies. What we are left with, instead, is technologies piling up for young, urban speakers of English (as well as a few other groups), increasing the inequality gap between them and (most of) the rest of the world.

The story does not stop here. Not all technologies are developed with Subgroup Test Ballooning. Sometimes technologies are developed for, what is believed to be representative samples of the current end user population, across, e.g., languages and demographic groups. This, at the face of it, sounds much fairer than Subgroup Test Ballooning but the two can be very hard to distinguish on markets dominated by young, urban speakers of English. Such **Snapshot-Representative Evaluation** of new technologies–representative only of the current snapshot of the end user population–calls for a slightly different response: The problem here is not the explicit test ballooning of a technology with a demographic subgroup but the assumption that we can and should sample from our current end users. Why is that a bad idea? First, end user populations tend to drift, for instance in the case of an expanding market. Second, we do not necessarily want to mirror the status quo. We often want to encourage drift, e.g., by obtaining gender balance, and put more weight on minority groups to mitigate data biases and induce fairer models. Subgroup Test Ballooning and Shap-Shot Representative Evaluation in tandem can reinforce existing inequalities, because subgroups that see better performance, will be more loyal end users. That is: Gaps in representation leads to gaps in performance, which in turn widen gaps in representation, leading to a vicious cycle.

Thought experiment: the egalitarian martian 2 The Martian evaluates speech recognition on Planet Earth and begins with the case of Danish. Danish speech recognition technology works better for young Danes than for old Danes, and as a consequence, young Earthlings use the technology more frequently. The Martian observes how the multinational technology companies – as well as the university research labs—that develop speech recognition models, optimize their performance on data collected from randomly selected users from their user pool. On average, this leads to a 4/5 over-representation of young users and a (too) homogeneous feedback signal. This, in turn, biases the model to do well on voices of young users, at the expense of older voices. Over time this creates a vicious cycle where out-of-target users (elderly) are underrepresented in feedback signals and among new users.

Thought experiment: the egalitarian martian 3 Having seen the downstream impact of Danish speech technology, the Martian reports back to Supreme Leader Xaroline and suggests that the Martians adopt an AI policy forcing

[1] Amazon's Alexa, for example, which was launched in 2014, is only available in eight languages at the time of writing this, eight years later. Google's Assistant, launched in 2016, is available in 12 languages.

[2] Anecdotally, we have seen up to 900% increases in error rates with available models, moving from product target group members to non-native speakers of dialect. Such performance is prohibitive of adoption, leaving groups for which speech recognition could be particularly useful, disillusioned about the technology.

[3] This practice is also common in academia, where annotation projects tend to recruit annotators in their 20s (university students or workers on crowd-sourcing platforms). The AI/NLP has seen occasional calls for including annotator demographics in data statements, e.g., [1], but few practitioners have followed suit.

companies to achieve equal (or ϵ-equal[4]) performance across salient sub-populations. Supreme Leader Xaroline replies that she wishes to use AI technologies to compensate for widespread dyslexia in the Martian population. She proposes a temporary policy guaranteeing that for now, all technologies should work significantly better for dyslexics than for other Martians.

In this article, we argue that Subgroup Test Ballooning and Snapshot-Representative Evaluation are unjust practices. How do such practices come about, and what motivates them? We argue—and this is the main contribution of our article—that there is a loophole for such practices in how most AI/NLP practitioners think about fairness. AI/NLP practitioners, as shown in the next section, rely on a Rawlsian conception of fairness. We show how Rawls' notion of fairness allows for Subgroup Test Ballooning and Snapshot-Representative Evaluation. We then compare Rawls' fairness to a more egalitarian notion of fairness. Such a notion of fairness has fewer loopholes. If AI/NLP is to avoid contributing to increasing global inequality, it should adopt a different definition of fairness prohibiting Subgroup Test Ballooning and Snapshot-Representative Evaluation.

2 AI/NLP fairness is Rawlsian

AI/NLP researchers have uniformly adopted a Rawlsian notion of fairness. This is reflected in the by now common practice of citing Rawls when mentioning fairness [3–7]. Fairness plays a central role in the philosophy of John Rawls. Social institutions must be fair to all members of society, regardless of background and dispositions. How, though, does he define fairness? This is seen from his theory of distributive justice, from *A Theory of Justice* (1971) [8], in which he writes:

> *Social and economic inequalities are to be arranged so that they are both:*
>
> (a) *to the greatest benefit of the least advantaged, consistent with the just savings principle, and*
> (b) *attached to offices and positions open to all under conditions of fair equality of opportunity.*

Principle (a) is often referred to as the Difference Principle and is the main focus of our discussion of Rawlsian fairness. Principle (a) does *not* enforce strict quality, but simply asks for the maximization of 'benefit of the least advantaged'. Benefit, for Rawls, is wealth or goods but in the context of AI/NLP we distribute performance. Rawls thus asks us to focus on raising the performance floor, rather than, say, minimizing the variance in performance across subgroups. Opinion divides on how much better off the least advantaged would be under the Difference Principle than under a strict equality principle. Rawls is not opposed to strict equality but is more concerned about the absolute position of the least advantaged group rather than their relative position. In Section 5, we will argue the relative position of the least advantaged is–or at least, can be–more important than the absolute position in the case of technologies such as AI/NLP.[5]

The algorithmic equivalent of Rawls' notion of fairness is 'maximizing the welfare of the worst-off group' [5]. A few things are left underspecified here. The first question, of course, is how to define groups. Groups are typically thought of as the product of a subset of protected attributes, e.g., gender and race.[6] Welfare, like 'benefit', is performance as measured by the go-to performance metric.[7] Rawlsian fairness thus becomes maximizing the performance on data sampled from the group on which performance is currently lowest. Many algorithms have therefore been developed to maximize performance on the groups with the worst performance.[8]

The AI/NLP literature does not compare Rawlsian fairness with alternative frameworks for thinking about fairness. There is considerable disagreement how best to quantify welfare [2, 16], i.e., what metrics to use, but not on the overall framework. What has also not been discussed in the literature, is the fact that Rawlsian fairness often tolerates considerable inequalities. We turn to a comparison of Rawlsian fairness with a more egalitarian alternative:

3 Rawls and Nielsen

Rawls' Difference Principle requires that economic systems be organized so that the least advantaged members of society are better off than they would be in any alternative economic

[4] Most practical fairness metrics measure approximate fairness by quantifying subgroup deviations [2]; subgroup performance is ϵ-equal or ϵ-fair if deviations are smaller than ϵ.

[5] We think this discussion is more constructive than the more general discussion of whether to slow down AI/NLP research and development, or opt for a more integrative approach [9]. As [9] note, the first option is not really on the table anyway.

[6] Such groups are sometimes referred to as *categories* in social science research [10].

[7] Most AI/NLP tasks come with multiple performance metrics, and it is often common practice to average across several metrics.

[8] Examples include square root sampling [11], adaptive scheduling [12], loss-balanced task weighting, [13], group-distributional robust optimization [14], and worst-case-aware automated curriculum learning [15].

arrangement. From the Difference Principle, we can derive what counts as justifications for inequality. Rawls' concern is about the absolute position of the least advantaged group rather than their relative position, and whether it is possible to raise the position of the least advantaged further, even at the cost of strict equality of income and wealth. If so, the Difference Principle prescribes inequality up to that point. The Difference Principle is, thus, in a sense, a loophole for inequalities. Rawls holds, for example, that inequalities that arise from our rewarding of acquired competencies under equal opportunity, are still fair, provided they make society richer or better.

Let us compare this with a more egalitarian definition of fairness, namely that of Kai Nielsen [17]. Nielsen's principle is a little different:

> *After provisions are made for common social (community) values, for capital overhead to preserve the society's productive capacity and allowances are made for differing unmanipulated needs and preferences, the income and wealth (the common stock of means) is to be so divided that each person will have a right to an equal share.*

The loopholes left open by this principle are fewer than with Rawls', allowing for only two exceptions to strict equality, namely what it takes to make basic services run (capital overhead), and to cover for people with special needs, e.g., impairments or illnesses.

So, the fairness principles of Rawls and Nielsen differ. Rawls allows for a higher degree of inequality and would argue that "an equal division of all primary goods is irrational in view of the possibility of bettering everyone's circumstances by accepting certain inequalities." This, of course, depends on your definition of what 'better' means, as discussed in length by Nielsen. We will contribute to this discussion in Sec. 5 but from the perspective of AI/NLP technologies, arguing that focusing exclusively on the absolute position of the worst off while allowing for significant performance disparities, is, in this context, a dangerous path to take. If our definition of fairness for AI/NLP is to prohibit Subgroup Test Ballooning and Snapshot-Representative Evaluation, Nielsen's definition of fairness seems more adequate than Rawls'.

4 AI/NLP loophole shooting

Early-stage development of technology focusing on available English benchmarks, and with an eye to technology-ready target audiences in rich countries, is common in AI/NLP. On Rawls' definition of fairness, such Subgroup Test Ballooning can be motivated by possible advancements bettering everyone's circumstances once technologies are transferred to other languages: Many AI/NLP papers on English claim that they "plan to scale to other languages" [18–21] but often never do. Some of the most popular benchmarks are know to exhibit demographic biases [22] but remain popular. Let us consider these justifications of AI/NLP-induced inequalities in more detail:

Justifications of inequalities Unfortunately, a large-scale empirical study of justification strategies in AI/NLP is yet to be undertaken but we briefly summarize a related study of justifications used in discussions of *income* equality [23]. The study finds five frames of justifications of inequality in discussions of income equality: (equal) opportunity, desert, procedure (of income determination), need, and (frame of) reference.[9] We present examples of what this could mean in an AI/NLP context, using Subgroup Test Ballooning on English as our example:

Justification	Frame
English is easy to learn; resources are abundant.	Opportunity
English is the most widely used language.	Desert
It is up to industry/research labs to decide.[10]	Procedure
English users have more advanced needs.	Need
Other technologies are for English markets first.	Reference

We have anecdotally come across all of the five frames in discussions in the AI/NLP community–and some have also surfaced in the academic literature [24–27]–but the list is likely incomplete, and the frames listed may differ significantly in popularity. This remains left for a more systematic study to decide.[11]

[9] That is: Income inequality is legitimate if everyone had (formally and substantively) equal opportunities to advantage (Opportunity); if everyone is compensated proportionally in terms of input (e.g., working time, education) and output (e.g., corporate success, social returns); if the inequality results from an agreed-upon established process (Procedure); if it reflects intrinsic or functional needs (Need); or if inequality is proportional to accepted standards in a particular domain (Reference).

[10] While intellectual freedom is crucial in AI/NLP fairness, this justification strategy also suggest a possible conflict: For, should researchers be free to pursue unfair technologies?

[11] One complicating factor is that some frames are used more explicitly than others. Opportunity arguments [24, 25] and Desert arguments [26, 27] are abundant in the academic literature, whereas you rarely see explicit Procedure, Need and Reference arguments, except for indirect Need arguments from researchers who are worried that AI/NLP researchers working on low-resource languages develop technologies for people who do not see the need for them. Our response to this form of justification of inequality would be to agree with the basic assumption that we are in no position to decide on behalf of people what technologies they ought to adopt. Our conclusion is different, however: If we are not to decide for people, we need to make technologies available to them. Otherwise we have decided on their behalf.

Thought experiment: the egalitarian martian 4 Our Martian field worker sees significant push back from his fellow Martians, who find his egalitarian proposal too radical. The push back has nothing to do with Supreme Leader Xaroline's correction, giving dyslexics special status. Neither is it because his fellow Martians worry egalitarian fairness will slow technological development. The push back comes from other government workers feeling egalitarian fairness is somehow unfair. The government workers reason as follows: If early adopters develop new habits, their technological maturity level increases, but egalitarian fairness will mean they have to wait for everyone to catch up, before they can enjoy new technologies. Question is whether their technological maturity justifies inequality? Our Martian field worker pushes back against this idea in a televised address to the nation: "Nothing in these policies prevent new technologies from being developed," he says, "as long as these technologies work equally good for all of us."

5 Why relative, not absolute position

We will present three arguments for why, in the context of AI/NLP performance across subgroups, the relative position of the least advantaged is more important than their absolute position:

1. *Staying On the Radar*: The absolute performance will improve rapidly for active end users but if we want to keep all subgroups represented among our end users, without anyone falling off our radar, we need to minimize the relative performance disparity across subgroups.
2. *Being Right for the Right Reasons*: High performance on data from some subgroups but high overall performance disparities, is typically sign of overfitting, i.e., reliance on spurious correlations in the data. Minimizing disparity across subgroups increases the likelihood of finding robust estimators, i.e., models that rely on factors that are robustly predictive.
3. *Breaking the Hype Cycle*: The absolute position of the most advantaged subgroup sets the expectations of everyone. The least advantaged will be more disillusioned with the technology, the larger the gap between them and best-case performance on the most advantaged subgroup.

Argument 1: staying on the radar Turn-around in AI/NLP is fast, and models quickly go from struggling on new benchmarks to surpassing human performance–a phenomenon known as *benchmark saturation* [28]. Benchmarks seem to saturate faster and faster and often within the first year or two of their publication.[12] Absolute performance is thus a rapidly moving target. The benefits of tolerating inequality in favor of higher absolute performance may, in other words, be short-lived. Also, tolerating performance gaps may create a vicious cycle. If a technology is clearly biased against your group, chance is you will abandon the technology. Your group will become under-represented in the pool of end users, performance on your group will not be optimized for and eventually deteriorate, making it less likely that your peers will choose to become users of the technology in question. Your subgroup falls of the technology's radar, so to speak.

Argument 2: being right for the right reasons Young and old speak slightly different languages but learning what groups have in common reduces the change of relying on spurious correlations. Consider the following examples of group-specific spurious correlations: (a) In movie review sentiment analysis, both young and old will speak of good and bad movies but groups may differ on whether they associate specific words such as *fast-paced* or *psychological* with positive or negative polarity. In reality, these words are not sentiment words but simply words that (within groups) covary with sentiment. Young people may associate the word *fast-paced* with positive sentiment but this predictor is not robust across groups. Systems that rely on such spurious correlations will be sensitive to drift in the user population, whereas a system that does not rely on such words – potentially compromising performance a bit – will be more robust. Such robustness is not just motivated by temporal drift but also the need to adopt to unseen product types and review platforms. b) In machine translation from English to German, reordering is sensitive to phrase boundaries. Punctuation is often a give-away for phrase boundaries but subgroups may differ in how consistently they use punctuation. Young people are, for example, less inclined to use punctuation in weblogs [29]. We know from psycho-linguistics that human sentence processing is *not* sensitive to punctuation, and in many domains, say in emails or on some social media, punctuation is almost entirely absent. It should thus be possible to infer phrase boundaries in the absence of punctuation, and a model that learns to do so, will obviously be more robust to variation.

Argument 3: breaking the hype cycle Our third argument for worrying more about the relative position of the least advantaged than their absolute position, has a more psychological flavor. In practice, technology development is often a matter of anticipating user disappointment. A

[12] The SuperGLUE Benchmark (https://super.gluebenchmark.com), for example, was launched on 6 May 2019 with a machine baseline and a human baseline 18.3% ahead of it, but was saturated by a newer model on 6 Jan 2021, surpassing human performance by half a percentage point.

machine translation model may err rarely but if it errs on translation problems that are considered obvious by most, end users will loose trust in the translation model. Since the first wave of early adopters of a new technology will be responsible for initial reviews and, possibly, early hype, they also set the expectations of later waves of users. Such users will likely be disillusioned if initial reviews and hype were based on overly optimistic performance estimates, because the technology was fitted on data sampled almost exclusively from the subgroup to which the early adopters belong. Voice assistants has seen performance gaps with some subgroups, e.g., females, low/middle-income, and low product experience, and some of these subgroups are also particularly sensitive to products' failure to meet their expectations [30]. Such dynamics easily create vicious cycles.

6 Fair systems and fair metrics

We have argued that Rawlsian fairness – widely adopted in AI/NLP research – admits for loopholes that are actively facilitating the development of biased technologies, including technologies biased toward certain languages and certain subgroups.[13] We have given examples of different types of justification of inequalities facilitated by such loopholes, and suggested a more egalitarian notion of fairness to prevent such practices. In general, we have argued that, in the context of AI/NLP model development, the relative position of the least advantaged may be more important than their absolute position. Finally, we want to emphasize that adopting a more egalitarian fairness principle will not 'solve' the fairness challenges in AI/NLP research once and for all.

Consider, for example, the role of performance metrics: Machine translation systems have to be fair with respect to dyslexia (a protected attribute). This, in our view, would mean that the performance of such a system, as measured through standard performance metrics such as BLEU, METEOR or something better,[14] must be equal for dyslexics and non-dyslexics. Of course output with equal BLEU scores need not be equally useful to two target groups, and in this case, it is plain to see that the machine translation system would have to produce somewhat-easy-to-read output to be useful to dyslexics. So, does this not show how a system with equal performance across groups can still be unfair?

We would argue that the limited usefulness of machine translation systems for dyslexics, is not a sign that these systems are as such unfair but that the evaluation metrics we use to evaluate them, are unfair. It is, in other words, unfair to dyslexics that readability is not part of how machine translation systems are evaluated. Clearly, adopting fairer metrics would have an impact on what models are induced but for now, models can be fair *with respect to* standard metrics without considering readability. We believe it is important to distinguish between model fairness and metric fairness to move research forward in the best possible way. AI/NLP models can also be unfair in other ways, e.g., protecting the privacy of end users using one operating system rather than another. We believe, however, that performance disparities across languages and groups are one of the most important roadblocks on the path toward fair AI/NLP models that do not widen existing inequality gaps between us.

7 Concluding remarks

Our argument in the above is simple: Rawlsian fairness–i.e., his Difference Principle–is too permissive to prevent common AI/NLP practices that actively contribute to global and social inequality gaps. Examples include test-ballooning technologies on specific target groups that are known to be adoption ready, or evaluation technologies on representative samples of the current end user population. We suggest a more egalitarian definition of fairness–adopted from Kai Nielsen's work on justice at large. We believe this will be an important step toward more sustainable AI/NLP research and development.

The trajectory of AI/NLP research and development is tied to its evaluation methodologies and performance metrics, and recent focus on fairness is an opportunity to course-correct for unjust practices by implementing less biased evaluation methodologies. Details matter, however, and now is a good time to get things right. If we want users and yet-to-become-users around the world to benefit equally from AI/NLP technologies, and if we want to avoid contributing to existing inequality gaps through unjust practices, we need to close the loopholes in current definitions of AI/NLP fairness.

Declarations

Conflict of interest The authors whose names are listed above certify that they have NO affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

[13] Fairness across languages and fairness across subgroups form a continuum because of dialects and sociolects. Somewhat ironically, the (demographic) AI/NLP fairness literature has been shown to be particularly biased toward English [31, 32].

[14] BLEU and METEOR are performance metrics based on the *n*-gram overlap between a system translation and one or more human reference translations. The metrics generally seem to correlate with human judgments of translation quality but they are considered far from perfect.

References

1. Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Trans Assoc Computational Linguist **6**, 587–604 (2018). https://doi.org/10.1162/tacl_a_00041
2. Williamson R., Menon A. Fairness risk measures. In: Chaudhuri K., Salakhutdinov R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6786–6797. PMLR, Long Beach, California (2019). https://proceedings.mlr.press/v97/williamson19a.html
3. Larson B. Gender as a variable in natural-language processing: Ethical considerations. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 1–11. Association for Computational Linguistics, Valencia, Spain (2017). https://doi.org/10.18653/v1/W17-1601. https://aclanthology.org/W17-1601
4. Vig J., Gehrmann S., Belinkov Y., Qian S., Nevo D., Singer Y., Shieber S. Investigating gender bias in language models using causal mediation analysis. In: Larochelle H., Ranzato M., Hadsell R., Balcan, M.F., Lin H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 12388–12401. Curran Associates, Inc., Vancouver, CA (2020). https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf
5. Ethayarajh K., Jurafsky D. Utility is in the eye of the user: A critique of NLP leaderboards. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4846–4853. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.393. https://aclanthology.org/2020.emnlp-main.393
6. Li M., Namkoong H., Xia S. Evaluating model performance under worst-case subpopulations. In: Ranzato M., Beygelzimer A., Dauphin Y., Liang P.S., Vaughan J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 17325–17334. Curran Associates, Inc., Vancouver, CA (2021). https://proceedings.neurips.cc/paper/2021/file/908075ea2c025c335f4865f7db427062-Paper.pdf
7. Chalkidis I., Pasini T., Zhang S., Tomada L., Schwemer S., Søgaard A. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4389–4406. Association for Computational Linguistics, Dublin, Ireland (2022). https://aclanthology.org/2022.acl-long.301
8. Rawls, J.: A Theory of Justice, 1st edn. Belknap Press of Harvard University Press, Cambridge, Massachussets (1971)
9. Cremer, D., Kasparov, G.: The ethics of technology innovation: a double-edged sword? AI Eth (2021). https://doi.org/10.1007/s43681-021-00103-x
10. Forsyth D.R.: Group Dynamics. Cengage Learning, Boston, MA (2009). https://books.google.dk/books?id=RsMNiobZojIC
11. Stickland A.C., Murray I.: BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In: ICML (2019)
12. Jean S., Firat O., Johnson M.: Adaptive scheduling for multi-task learning. In: ArXiv 1909.06434 (2019)
13. Liu S., Liang Y., Gitter A.: Loss-balanced task weighting to reduce negative transfer in multi-task learning. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 9977–9978 (2019). https://doi.org/10.1609/aaai.v33i01.33019977
14. Sagawa S., Koh P.W., Hashimoto T.B., Liang P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
15. Zhang S., Zhang X., Zhang W., Søgaard A.: Worst-Case-Aware Curriculum Learning for Zero and Few Shot Transfer. arXiv (2020). https://doi.org/10.48550/ARXIV.2009.11138. https://arxiv.org/abs/2009.11138
16. Hedden, B.: On statistical criteria of algorithmic fairness. Philos Pub Aff **49**(2), 209–231 (2021). https://doi.org/10.1111/papa.12189
17. Nielsen, K.: Radical egalitarian justice: justice as equality. Soc Theory Pract. **5**(2), 209–226 (1979)
18. Antworth, E.L.: Book reviews: Computational morphology: Practical mechanisms for the English lexicon. Computational Linguistics. (1992)
19. Tosik M., Lygteskov Hansen C., Goossen G., Rotaru, M.: Word embeddings vs word types for sequence labeling: the curious case of CV parsing. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 123–128. Association for Computational Linguistics, Denver, Colorado (2015). https://doi.org/10.3115/v1/W15-1517. https://aclanthology.org/W15-1517
20. Vylomova E., Cotterell R., Baldwin T., Cohn, T.: Context-aware prediction of derivational word-forms. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers, pp. 118–124. Association for Computational Linguistics, Valencia, Spain (2017). https://aclanthology.org/E17-2019
21. van Erp M., Groth P.: Towards entity spaces. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2129–2137. European Language Resources Association, Marseille, France (2020). https://aclanthology.org/2020.lrec-1.261
22. Hovy D., Søgaard A.: Tagging performance correlates with author age. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 2: Short Papers), pp. 483–488. Association for Computational Linguistics, Beijing, China (2015). https://doi.org/10.3115/v1/P15-2079. https://www.aclweb.org/anthology/P15-2079
23. Bank, J.: Mr. winterkorn's pay: A typology of justification patterns of income inequality. Soc Justice Res **29**(2), 228–252 (2016)
24. Utiyama M., Isahara H.: A comparison of pivot methods for phrase-based statistical machine translation. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 484–491. Association for Computational Linguistics, Rochester, New York (2007). https://aclanthology.org/N07-1061
25. Anastasopoulos A., Lui A., Nguyen T.Q., Chiang D.: Neural machine translation of text from non-native speakers. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), pp. 3070–3080.

Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1311. https://aclanthology.org/N19-1311
26. Blasi D., Anastasopoulos A., Neubig G.: Systematic inequalities in language technology performance across the world's languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), pp. 5486–5505. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.376. https://aclanthology.org/2022.acl-long.376
27. Lewis P., Oguz B., Rinott R., Riedel S., Schwenk H.: MLQA: Evaluating cross-lingual extractive question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7315–7330. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.653. https://aclanthology.org/2020.acl-main.653
28. Kiela D., Bartolo M., Nie Y., Kaushik D., Geiger A., Wu Z., Vidgen B., Prasad G., Singh A., Ringshia P., Ma Z., Thrush T., Riedel S., Waseem Z., Stenetorp P., Jia R., Bansal M., Potts C., Williams A.: Dynabench: Rethinking benchmarking in NLP. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4110–4124. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.324. https://aclanthology.org/2021.naacl-main.324
29. Burger J., Henderson J.: An exploration of observable features related to blogger age. In: AAAI 2006 Spring Symposia, pp. 15–20 (2006)
30. Brill, T., Munoz, L., Miller, R.: Siri, alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. J Mark Manag (2019). https://doi.org/10.1080/0267257X.2019.1687571
31. Takeshita M., Katsumata Y., Rzepka R., Araki K.: Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 44–55. Association for Computational Linguistics, Barcelona, Spain (Online) (2020). https://aclanthology.org/2020.gebnlp-1.5
32. Ruder S., Vulić I., Søgaard A.: Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 2340–2354. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.findings-acl.184. aclanthology.org/2022.findings-acl.184