Artificial intelligence has made unprecedented progress in language applications. Nonetheless, there still remains a large gap between human and machine understanding of language, in all of its complexities. A primary driver of this discrepancy is the disproportionate focus on increasing *quantities* of training data, at the cost of data *quality*. This trend persists even as the community recognizes that our datasets abound with undesirable biases—both unintended correlations and more pernicious social biases, perpetuated and exacerbated by models. Concurrently, large models have become pivotal in AI, but often ignore prior information about the structure of the task (e.g., world or domain knowledge) and its inputs (e.g., linguistic structure) which can offer valuable signals, or inductive biases for training. My research seeks to build robust models of language by **consolidating the potent inductive biases** into the models, while **addressing the undesirable biases** in our existing, albeit flawed, training datasets. Towards this goal, I have developed models and learning algorithms spanning three research directions:

Algorithmic Discovery and Mitigation of Dataset Biases (§1): I have demonstrated the presence of both spurious and social biases in well-known datasets via a suite of tools including adversarial algorithms, as well as proposed mitigation strategies for these biases [1, 2, 3, 4, 5].

Inductive Biases from Explicit Structural Scaffolds (§2): I have explicitly incorporated inductive biases from linguistic structures, via learning objectives and algorithms; these act as scaffolds supporting robust representation learning in various semantic tasks involving structured prediction [6, 7, 8, 9, 10, 11].

Inductive Biases from Implicit Priors (§3): My research has integrated inductive biases from implicit prior knowledge, such as the textual domain and the structure of tasks, showing room for improvement over language representations learnt at scale [12, 13, 14, 15].
I describe each direction in detail below, and conclude with challenges I aim to explore in future work (§4).

1 Algorithmic Discovery and Mitigation of Dataset Biases

The advent of crowdsourcing platforms has enabled collection of large labeled AI datasets. However, crowd annotators often follow labeling heuristics for speed, diluting the data quality. Consider the Stanford Natural Language Inference dataset [16, SNLI], which labels if a textual hypothesis is true given a premise. In our work at NAACL 2018 [4], we showed that SNLI contained many annotation artifacts, which lacked any grounding in semantics. For instance, most negation words, and surprisingly the word 'cat' in the hypothesis were indicative of a contradiction, *regardless of the premise*, simply due to co-occurrence. In many similar settings, models tend to rely on such spurious dataset biases, limiting their ability to generalize to out-of-distribution examples. My research has proposed algorithms [1, 2] and architectures [3, 4] aimed at improving model generalizability through the discovery, and mitigation of such biases.

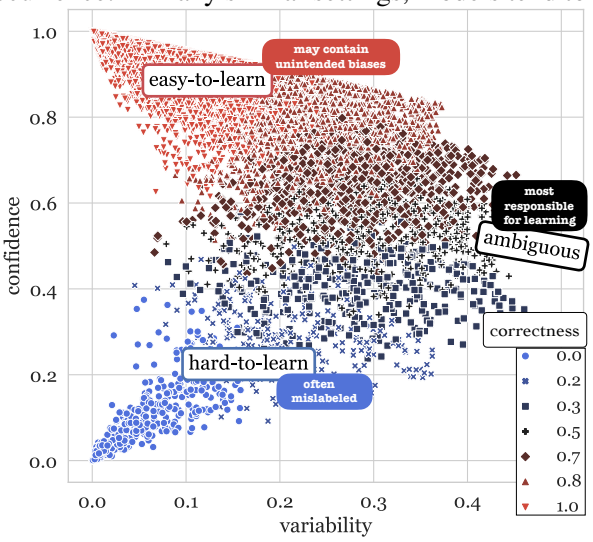

Figure 1: Data Map showing SNLI instances, with respect to a Transformer model. Distinct regions are labeled.

Bias Discovery A key intuition for data bias discovery comes from the differential contribution of training instances towards learning. Our EMNLP 2020 paper on Dataset Cartography [1] introduces **Data Maps** (see Fig. 1) to contextualize each data instance based on its behavior during training. Training instances which are predicted correctly (high confidence) and consistently (low variability) throughout training are easy to learn for the model. Instances which the model consistently predicts incorrectly, are, analogously, hard to learn. We showed that easy-to-learn examples are disproportionately associated with unintended biases [5], while hard-to-learn instances are frequently mislabeled. Training instances on which the model predictions are inconsistent comprise a third category: ambiguous, which are key for performance, both in and out of distribution. Our method discovers dataset regions based on training behavior (training dynamics), as a by-product of training. This could allow dataset collectors to limit spurious biases that unintentionally steer the model away from true generalization.

As mentioned above, instances which are *easy* for a model, are often so due to the presence of spurious biases. Our work at ICML 2020 [2] describes an algorithmic approach to discover such instances, AFLite (Lightweight

Adversarial Filters). This algorithm determines the *predictability* of a data instance based on the consensus of an ensemble of simple linear classifiers. It iteratively filters out instances which are deemed too predictable, thereby selecting progressively adversarial subsets of the data for a model to train on. We showed that AFLite is a greedy approximation of an optimal, yet intractable bias reduction algorithm. Results on *AFLite-filtered test sets* revealed that the performance of state-of-the-art models is often inflated, and contrary to previous findings, is still behind human performance, due to model reliance on spurious biases. Thus, AFLite-filtered test sets are more accurate benchmarks.

Simplification of model architectures can also identify instances with spurious biases. Models which perform well despite receiving only part of the input (e.g. only the hypothesis in SNLI) can reveal biases [4]. Others, such as our approach from ACL 2020 [3], involve architectures which permit early exits, i.e. allow inference to proceed with only a few layers of a deep neural stack. We showed that instances which are accurately predicted despite early exits are likely to contain spurious biases.

Bias Mitigation Robust models of language will generalize to real world settings, and therefore be resilient to spurious biases in datasets. Removal of predictable (for AFLite [2]) or easy-to-learn instances (for Data Maps [1]) from data presents an intuitive bias mitigation strategy. Remarkably, in the absence of these instances in the training data, models were able to *generalize better to challenging, out-of-distribution test sets* on a variety of NLP tasks, even though trained on a fraction of the data.

However, mitigation of **social biases** in datasets still remains a lofty goal. Datasets for online hate speech detection, for instance, are known to contain pernicious racial and dialectal biases, directed towards minorities like speakers of African American English [17]. These biases are only exacerbated by models, due to spurious, lexical and dialectal associations. In our current work under review [5], we study the efficacy of AFLite, Data Maps and auxiliary training objectives from prior work for social bias mitigation. While all three methods are effective to some extent, we found that resultant models still exhibit racial and dialectal biases. Further investigation indicated that addressing social biases in datasets might only be possible through meticulous relabeling, with special attention to annotators' implicit predispositions towards social groups. In the long run, dataset bias mitigation and removal remains an important open problem.

2 Inductive Biases from Explicit Structural Scaffolds

At the heart of AI lies language understanding, or the communication of the speaker intent to the listener. Linguistic structure given by syntactic trees (see Fig. 2) provide strong priors for language understanding in humans [18]. These priors, when explicitly incorporated in models, serve as **inductive biases**, or signals that help the model make the right decision. Thus, in contrast to spurious dataset biases, linguistically-motivated model inductive biases can yield rich language representations, robust to the spurious biases. My research has focused on linguistic structural scaffolds for robust language representations via learning objectives and algorithms.

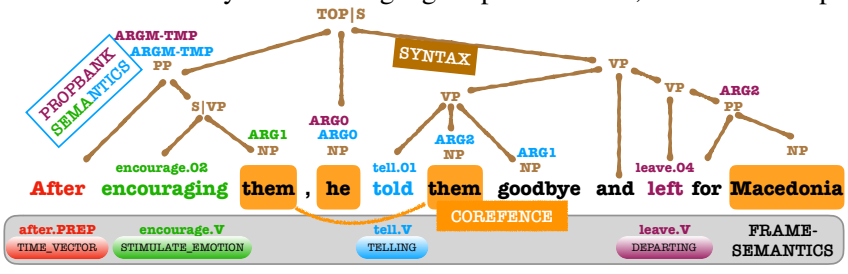

Figure 2: Shared structural similarities between various semantic and syntactic formalisms. Semantic arguments often overlap with syntactic constituents, and with one another, across formalisms. Syntactic constituency tree (Above; Brown), semantic predicates and arguments (Top; Color-Coded), coreference chains (Overlay; Orange) and semantic frames [19] (Below; Color-coded; roles not shown) for a sentence.

Syntactic Structural Scaffolds Semantics is built on the foundation of syntax; words and phrases which fill certain syntactic roles are also likely to fill semantic roles. For example, in Fig. 2, the phrase *'After encouraging them'* is both the syntactic descendant of the verb *'told'*, as well as a semantic descendant of the 'Telling' event; access to the syntactic structure can help learn the semantics. We leverage this intuition for jointly learning syntactic trees with semantic graphs in our CoNLL paper [6]. We construct a (multi-)graph for both structures with distinct edges for syntactic and semantic relations, and nodes for words in the sentence. Our greedy linear-time algorithm predicts the graph, one edge at a time, tracked using stack data structures. Our joint approach proved more accurate than a syntax-agnostic baseline, highlighting that syntax indeed helped learn semantics.

Learning the complete syntactic tree for a sentence, however, can be expensive. Moreover, only a subset of syntactic structure overlaps with semantic arguments. In our work on Syntactic Scaffolds at EMNLP 2018

[7], we showed that even using a **shallow approximation** of relevant syntactic structure was beneficial. We showed this for three semantic tasks, across diverse formalisms of meaning, each classifying spans of text as component parts of a semantic graph (see an example in Figure 2). By bypassing full syntactic parsing, this partial scaffold promoted efficiency both during training and testing, as well as improved performance over non-syntactic baselines and prior work [8]. Moreover, our setting avoided the necessity for both syntactic and semantic supervision *on the same data*; structural scaffolds can thus be applied to a variety of other tasks.

Similar to the syntactic-semantic symbiosis, various semantic formalisms can have complimentarity benefits. In [9], my collaborators and I observed gains from jointly learning two structurally divergent semantic formalisms—semantic dependency graphs and span-based semantic graphs. In [10], we found multilingual learning of semantic dependencies *across seven languages* beneficial for the resource-poor languages [20].

Given the usefulness of syntax, we asked whether large language models (LMs) could further benefit from partial syntactic scaffolds [11]. We proposed a hierarchical syntactic encoder (see Fig. 3), whose internal states captured the inductive biases from the shallow, predicted syntax. Downstream task performance did not show significant benefits from this syntactic encoder, indicating that LMs pretrained on large corpora *already capture* predicted, shallow syntactic information; this has implications for the learnability of syntax from weakly-biased models.

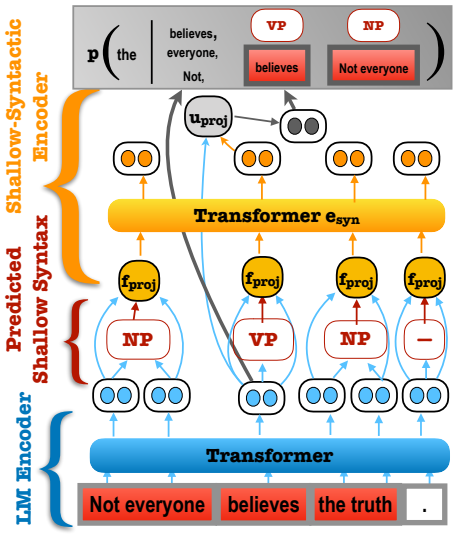

Figure 3: Hierarchical encoder for a shallow syntax-aware language model.

Learning Objectives for Scaffolds Objectives for *transferring* structural inductive biases to other tasks, via structural scaffolds, need careful design. This transfer learning [21] can be done via pipelining multiple objectives [11], joint training [6], multitask training with auxiliary objectives [7], or through marginalization of latent structure variables [14]. When the desired structures are sufficiently disparate, or differ in scales of supervision, such as large LMs and shallow syntax [11], training one objective after the other is effective. However, such staged training tends to cascade errors down the pipeline. Jointly training a single objective for syntax and semantics serves as an alternative to pipelines [6, 10], where the uncertainty of a syntactic (or semantic) decision is captured in intermediate joint structural representations. However, joint models require complex inference algorithms [6] to account for divergent structural constraints. A simplification is offered by learning multiple task objectives [9], where most parameters are shared between tasks. This framework is particularly useful when one (auxillary) task only serves to aid the primary task, for e.g., shallow syntax to aid full semantic structured prediction [7].

3 Inductive Biases from Implicit Priors

While explicit linguistic structural scaffolds are valuable inductive biases, they can be expensive to obtain and incorporate at scale. As a result, modern large-scale NLP vastly ignores prior knowledge for tasks. However, there exist easily available implicit, domain-specific priors, which can be incorporated at scale via exposure to even more, relevant data [12, 13], or task-specific knowledge [14, 15]. One focal point of my research building robust task- and domain-specific language representations using inductive biases arising from such priors.

Data Augmentation Today's NLP is built on language models pretrained on text from a variety of sources [22]. These sources, though very large, are finite and cannot capture the entire diversity of (the English) language. Our recent work at ACL 2020 [12] uses an easily available implicit prior—we continue to train the pretrained LMs on more unlabeled data in the target domain (**domain-adaptive pretraining**) before training on the task supervision. This straightforward inductive bias, which effectively augments the training data the model is exposed to, led to improvements in several domains, under both high- and low-resource settings. We additionally found that the data for augmentation can be obtained using a domain similarity-based data selection strategy, helpful in low-resource domains (e.g. scientific data), or when the domain is hard to ascertain (e.g. online data).

In contrast to unlabeled data, our work at EMNLP 2020 [13] obtained labeled data for augmentation cheaply via automatic generation with powerful LMs trained on the domain of interest. This provided useful inductive biases for commonsense reasoning tasks, where manual curation of training data is expensive [23], not to mention prone to spurious biases from human annotation [4]. Our method proved quite effective in low-resource settings [20], establishing a new state-of-the-art on various commonsense reasoning benchmarks. Post generation, we

selected the most informative and diverse generations (see Fig. 4) for training. This enhanced out-of-distribution generalization, proving more robust against adversaries or perturbations.

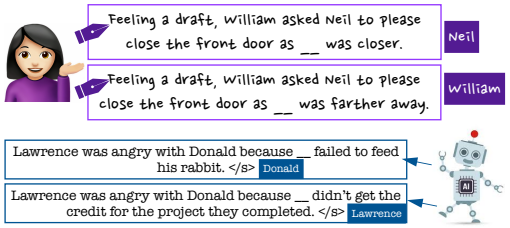

Figure 4: Human-authored vs. generated Winograd schema [23] twins for commonsense reasoning.

Task-Specific Priors Pretrained LMs [22] have allowed the use of off-the-shelf classifiers, ignoring task-specific priors. Consider question answering with distant supervision where the candidate answers for a question might occur anywhere in a lengthy web document (e.g. Wikipedia page) or might be mentioned in multiple contexts. However, knowing that the document is structured into pre-defined passages, and that answers occur in the same context as the question, offers useful inductive biases implicit in the task. In our work at ICLR 2018 [14], we designed a powerful latent variable model to incorporate these task-specific inductive biases. These priors allowed even a simple bag-of-embeddings architecture to surpass the performance of more powerful recurrent neural models, and improved space and time complexity.

Building a syntactic parser for tweets, without considering the implicit structure, can similarly be ineffective. Our work at EMNLP 2014 introduced TweeboParser [15] that de-selected syntactically irrelevant tokens (e.g. hashtags, emojis, hyperlinks) but considered multi-word expressions and multiple roots, atypical in standard English but common in tweets; both task-specific biases yielded gains over strong yet task-agnostic parsers.

4 Future Research Plans

My long term goals include developing models and algorithms for **equitable AI** that promote desirable biases, and address undesirable biases in datasets and models, with a focus on four specific directions:

Interpreting Biases The ability to interpret model decisions, and by extension, its underlying biases, is imperative for building trustworthy AI. Together with my research interns at AI2, I'm currently exploring algorithms to uncover the most salient signals in the input which influence a model's prediction. Information-theoretic methods, such as information bottleneck, aided by sparse structured priors, are well-suited to answer this question in an unsupervised setting [24]. Input attribution methods systematically reduce the input signal—a reduction with the largest change in model decision is thus the most influential. We use such attribution methods to explain decisions of models trained with structured and unstructured labels, key for social bias detection [25]. Explaining why a model made a certain decision, *in contrast to* another, yields a finer-grained AI interpretation; we explore such contrastive explanations via nullspace projections in the latent space to remove confounding factors [26].

Unlearning Social Biases Today's powerful pretrained language models are trained on enormous quantities of text from the web. But this data also contains undesirable social biases [5] in the form of significant amounts of toxic language, reflected, if not exacerbated in the language models. Is it possible to unlearn social biases in language models? Following up on my previous work on adapting language models *towards* target domains [12], I am interested in exploring if we can adapt pretrained models *away* from known sources of toxic language [27]. As part of my larger goal of social bias mitigation in AI (§1), this would involve careful design of negative training objectives, such that the expressiveness of the learnt representations remains unaffected otherwise.

Building Dynamic Benchmarks With the growing awareness of bias in datasets [1, 2], static datasets are no longer sustainable. Making datasets dynamic via iterative elimination of instances exhibiting undesirable biases, as well as addition of instances offering useful inductive biases from complex semantic and pragmatic reasoning, might be a promising alternative. Moreover, such benchmarks could promote generalization, in contrast to overfitting to specific trends in the static dataset. A new pool of data instances can be ranked automatically via measures from training dynamics [1], determining which instances still pose a challenge for state-of-the-art models. I am interested in building frameworks for dynamically growing datasets, and therefore dynamic evaluation benchmarks that evolve as models grow more powerful.

Contextualizing AI Predictions Given that humans use language to establish connections [28], record and disseminate information, language might always reflect human biases. What can the linguistic training data, metadata (e.g. annotators' identities), and algorithms then reveal about the appropriate use cases for a trained model? *In addition to* dataset bias mitigation, this information can help contextualize the ground truth [29] and therefore model predictions, with respect to social biases. I am excited to foster collaborations with theorists and sociolinguists to build formal contextualization frameworks for fair and equitable AI.

References

[1] **Swayamdipta, Swabha**, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics," in *Proceedings of ACL*, 2020. https://www.aclweb.org/anthology/2020.emnlp-main.746.

[2] R. L. Bras, **Swayamdipta, Swabha**, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi, "Adversarial filters of dataset biases," in *Proceedings of ICML*, 2020. https://icml.cc/virtual/2020/poster/6628.

[3] R. Schwartz, G. Stanovsky, **Swayamdipta, Swabha**, J. Dodge, and N. A. Smith, "The right tool for the job: Matching model and instance complexities," in *Proceedings of ACL*, 2020. https://www.aclweb.org/anthology/2020.acl-main.593.

[4] S. Gururangan, **Swayamdipta, Swabha**, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proceedings of NAACL*, 2018. https://www.aclweb.org/anthology/N18-2017/.

[5] X. Zhou, M. Sap, **Swayamdipta, Swabha**, Y. Choi, and N. A. Smith, "The ineffectiveness of algorithmic debiasing for toxic language detection," 2020. In Submission.

[6] **Swabha Swayamdipta**, M. Ballesteros, C. Dyer, and N. A. Smith, "Greedy, joint syntactic-semantic parsing with Stack LSTMs," in *Proceedings of CoNLL*, 2016. https://www.aclweb.org/anthology/K16-1019/.

[7] **Swayamdipta, Swabha**, S. Thomson, K. Lee, L. Zettlemoyer, C. Dyer, and N. A. Smith, "Syntactic scaffolds for semantic structures," in *Proceedings of EMNLP*, 2018. https://www.aclweb.org/anthology/D18-1412/.

[8] **Swayamdipta, Swabha**, S. Thomson, C. Dyer, and N. A. Smith, "Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold," 2017. arXiv preprint arXiv:1706.09528 https://arxiv.org/abs/1706.09528.

[9] H. Peng, S. Thomson, **Swayamdipta, Swabha**, and N. A. Smith, "Learning joint semantic parsers from disjoint data," in *Proceedings of NAACL*, 2018. https://www.aclweb.org/anthology/N18-1135/.

[10] P. Mulcaire, **Swayamdipta, Swabha**, and N. Smith, "Polyglot semantic role labeling," in *Proceedings of ACL*, 2018. https://www.aclweb.org/anthology/P18-2106/.

[11] **Swayamdipta, Swabha**, M. Peters, B. Roof, C. Dyer, and N. A. Smith, "Shallow syntax in deep water," 2019. arXiv preprint arXiv:1908.11047 https://arxiv.org/abs/1908.11047.

[12] S. Gururangan, A. Marasović, **Swayamdipta, Swabha**, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of ACL*, 2020. https://www.aclweb.org/anthology/2020.acl-main.740.

[13] Y. Yang, C. Malaviya, J. Fernandez, **Swayamdipta, Swabha**, R. L. Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "Generative data augmentation for commonsense reasoning," in *Findings of EMNLP*, 2020. https://www.aclweb.org/anthology/2020.findings-emnlp.90/.

[14] **Swayamdipta, Swabha**, A. P. Parikh, and T. Kwiatkowski, "Multi-mention learning for reading comprehension with neural cascades," in *Proceedings of ICLR*, 2018. https://openreview.net/forum?id=HyRnez-RW.

[15] L. Kong, N. Schneider, **Swayamdipta, Swabha**, A. Bhatia, C. Dyer, and N. A. Smith, "A Dependency Parser for Tweets," in *Proceedings of EMNLP*, 2014. "https://www.aclweb.org/anthology/D14-1108".

[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of EMNLP*, 2015. https://www.aclweb.org/anthology/D15-1075.

[17] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of ACL*, 2019. https://www.aclweb.org/anthology/P19-1163.

[18] N. Chomsky, *Syntactic structures*. Walter de Gruyter, 1958. https://psycnet.apa.org/record/1958-05423-000.

[19] C. F. Baker, M. Ellsworth, M. R. L. Petruck, and **Swayamdipta, Swabha**, "Frame semantics across languages: Towards a multilingual FrameNet," in *Proc. of COLING: Tutorials*, 2018. https://www.aclweb.org/anthology/C18-3003.

[20] C. Cherry, G. Durrett, G. Foster, R. Haffari, S. Khadivi, N. Peng, X. Ren, and **Swayamdipta, S.**, eds., *Workshop Proceedings: Deep Learning Approaches for Low-Resource NLP (DeepLo)*, 2019. https://www.aclweb.org/anthology/D19-6100.

[21] S. Ruder, M. E. Peters, **Swayamdipta, S.**, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of NAACL: Tutorials*, 2019. https://www.aclweb.org/anthology/N19-5004.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. https://arxiv.org/abs/1907.11692.

[23] H. J. Levesque, E. Davis, and L. Morgenstern, "The Winograd schema challenge," in *Proceedings of KR*, pp. 552–561, 2012. https://dl.acm.org/doi/abs/10.5555/3031843.3031909.

[24] J. Cao and **Swayamdipta, Swabha**, "Structured sparsity and an information bottleneck for extractive rationales," 2020. In Prep.

[25] J. Liang, **Swayamdipta, Swabha**, and C. Bhagavatula, "The effect of structured labels in interpreting social biases," 2020. In Prep.

[26] A. Jacovi and **Swayamdipta, Swabha**, "Explanations via contrasts," 2020. In Prep.

[27] **Swayamdipta, Swabha**, A. Liu, and Y. Choi, "Unadapting language models to biases," 2020. In Prep.

[28] **S. Swayamdipta** and O. Rambow, "The pursuit of power and its manifestation in written dialog," in *Proceedings of ICSC at IEEE*, 2012. https://ieeexplore.ieee.org/abstract/document/6337078.

[29] D. Raji, A. Paullada, S. Gururangan, A. Birhane, and **Swayamdipta, Swabha**, "Whose perspective is it anyway? Examining the Universality of Ground Truth via the Analysis of Toxic Language," 2020. In Prep.