The overarching aim of my research is to **make natural language processing (NLP) systems human-centric, socially aware, and equity driven**. Achieving this ambitious goal requires rethinking every stage of the machine learning pipeline by incorporating social awareness into data sets, training paradigms, and inference algorithms. Concretely, this requires machines to **understand the social dynamics implied in statements**, sentences in stories, or commands issued to an artificial intelligence (AI) assistants. For example, when a user asks to "*turn on the security cameras*," an AI assistant should be capable of inferring the underlying intents and reactions of the user–that the user "*is worried*" and "*wants to protect their property*". However, if the user specifies to turn the cameras on "*because someone with a headscarf just walked in*," this statement now also evokes the harmful implication that "*people with headscarves are seen as threatening*." An AI assistant should be able to understand and anticipate such harmful implications, e.g., to avoid exacerbating negative stereotypes about minorities.

During my PhD, I have taken several steps towards this ultimate goal of socially aware NLP systems. I will continue these efforts by focusing on three directions:

- **Interpersonal Commonsense Reasoning for Human-Centric NLP** – modelling social commonsense reasoning with machines and creating evaluation benchmarks to quantify reasoning abilities [1, 2, 3, 4, 5, 6, 7]
- **Representation Learning and Algorithms for Mitigating Social Biases** – distilling complex social and power dynamics between demographic groups into structured representations and incorporating those into neural models for revising text [8, 9, 10]
- **Diagnosing the Ethics and Fairness of NLP systems** – uncovering social biases in NLP systems and analyzing their provenance [11, 12, 13]

1 Interpersonal Commonsense Reasoning for Human-Centric NLP

Though trivial for humans, commonsense reasoning has remained an elusive goal for AI systems [14]. For example, given an event like "*X repels Y's attack*," we can easily make inferences about the event's likely causes ("*X wanted to save themselves*") and effects ("*X feels tired*" or "*Y will fall back*") even if we have never personally repelled someone's attack. Models struggle with this type of reasoning partly because they are trained on data that is inherently limited in coverage [e.g., due to reporting bias; 15], and partly because their neural architectures lead them to learn surface correlations instead of reasoning [16, 17].

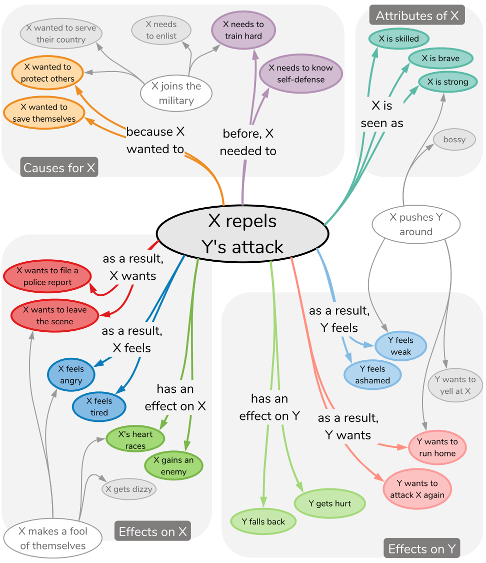

Figure 1: ATOMIC is a large knowledge graph that distills inferential knowledge about the causes and effects of events such as "X repels Y's attack".

As a step towards **modelling commonsense reasoning with machines**, I introduced ATOMIC [1], a large knowledge graph of inferential knowledge about the causes and effects of everyday situations represented in short natural language phrases (Figure 1). To achieve high coverage and large scale, we collected ATOMIC using crowdsourcing, to overcome reporting biases that prevent the use of automatic extraction methods for extracting this type of commonsense knowledge. Enabled by the large scale of natural language knowledge in ATOMIC, we developed COMET [3],[1] an inference engine that creates **neural representations of commonsense using pretrained language models**. With the knowledge

[1] https://mosaickg.apps.allenai.org/comet_atomic

learned during pretraining, COMET's inference ability generalizes to unseen events significantly better than non-pretrained baselines, as measured by automatic and human evaluations.

Another step towards achieving commonsense reasoning with machines involves evaluating the types of knowledge and reasoning that machines can do. Recently, I introduced **SOCIAL IQA** [2], the **first large scale social commonsense benchmark** to assess a model's ability to reason about the motivations, reactions, causes, and effects of social interactions in a question-answering (QA) format. For example, after "*Alex spills food on the floor, making a huge mess*," a model must choose whether Alex is more likely to "*taste the food*" or "*mop up*." To make this benchmark robust and avoid spurious correlations, we designed an adversarial crowdsourcing setup that collects unlikely answers using different but related questions, and filtered the data to remove instances with surface patterns [17]. This benchmark remains a challenge for state-of-the-art models (e.g., GPT-3 175B achieves 55% in a few-shot setting), partly because they are unable to distinguish participants or disentangle causes from effects.

Additionally, my colleagues and I introduced STORYCOMMONSENSE [5], a benchmark that focuses on model's ability to **infer the motivations and reactions of participants in the context of short stories**. Here, we framed the task as a set of multi-class classification tasks, where categories are grounded in psychological theories of motivations and emotions (e.g., Maslow's hierarchy of needs). Crucially, this benchmark requires reasoning about the mental states of characters even if they are not mentioned, and in the full past context of a story.

Future work Moving forward, my research will give models the ability to reason properly about social interactions. Specifically, I am excited to investigate novel neural architectures that do not rely merely on surface patterns or spurious correlations. For example, I am currently investigating model architectures that can track the mental states of participants in situations, to overcome the lack of partipant-centric reasoning that models exhibit on SOCIAL IQA [2]. I also want to continue drawing connections between social commonsense reasoning in AI systems and human cognitive processes, following my recent work investigating the types of knowledge included in stories drawn from episodic and semantic memory [7]. Additionally, I want to investigate new learning paradigms that incorporate causality and inferential reasoning, using neurobiologically plausible mechanisms inspired by neuroscience and cognitive science [e.g., predictive coding, Hopfield networks; 18, 19]. Finally, I look forward to adapting these commonsense reasoning abilities to conversational and therapeutic settings, drawing, for example, on my experience with building the chatbot that won the 2017 Alexa Prize [6].

2 Representation Learning and Algorithms for Mitigating Social Biases

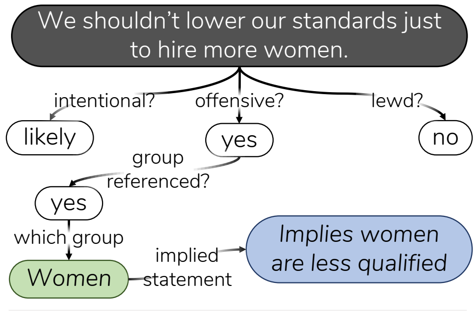

Figure 2: SOCIAL BIAS FRAMES is a new formalism for structured understanding of biased implications of text.

For machine learning systems to serve members of society equitably, they need to have a way to represent and account for social and power dynamics. The key challenge is that these dynamics are complex, and often not stated explicitly in data (e.g., "we shouldn't lower our standards just to hire more women" carries the subtle biased implication that "women candidates are less qualified"). Since biased language detection had been simplified to opaque yes/no decisions of offensiveness by AI systems, my work has tackled this challenge by distilling these complex social dynamics in novel structured representations. Additionally, I have tackled using these representations to help revise and debias textual data.

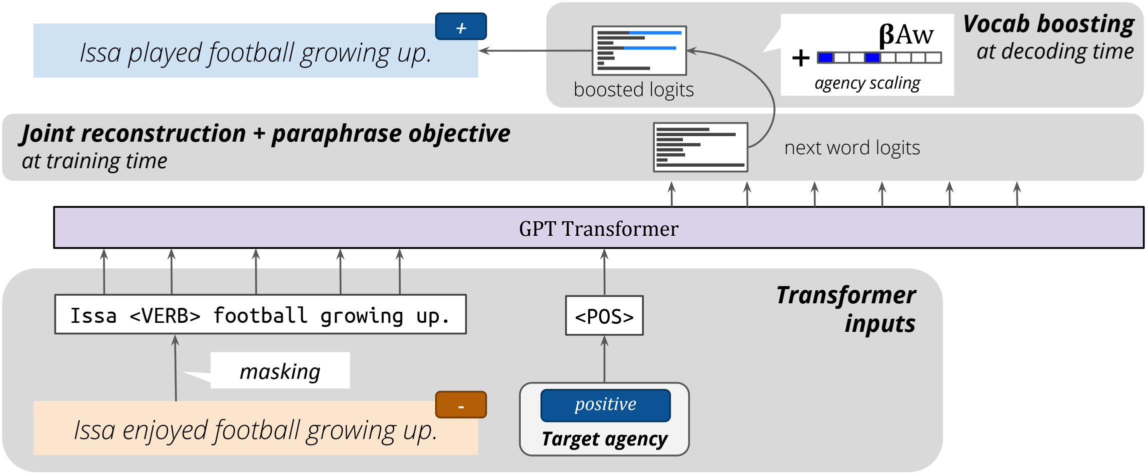

Figure 3: Overview of the PowerTransformer model for unsupervised controllable text revision, trained to debias the portrayal of characters in sentences.

I introduced **Social Bias Frames** [8], a **new formalism for representing biased implications and stereotypes** that are evoked in text (Figure 2). I designed the Social Bias Frame to distill knowledge into categorical variables of offensiveness and intent, paired with free text explanations of the targeted group and implied statement, to enable more explainable and trustworthy AI systems. Using a large corpus of social media posts annotated with our frames, we showed that neural models predicted offensiveness relatively easily but struggled to spell out subtle implications in text. This work won the best paper award at WeCNLP 2020.

As a first step towards debiasing text, I created **PowerTransformer** [10], a **new model for unsupervised controllable revision** that alters the portrayal of characters in story sentences. We debias portrayals through the lens of **connotation frames of power and agency** [9], a formalism that encodes pragmatic knowledge of implied power and agency dynamics with respect to predicates (e.g., “X plays football” portrays X as high agency, active, and decisive). In addition to the gender bias we uncovered in modern movies, subsequent work has used our connotation frames to study biases in news [20, 21], blogs [22], and school textbooks [23].

As illustrated in Figure 3, PowerTransformer was trained using a self-supervised denoising objective and an auxiliary paraphrasing objective, overcoming the lack of parallel training data for controllable revision tasks. Through ablation studies and human evaluation, we showed that our model benefits from both objectives independently and outperforms existing text revision baselines. Importantly, we used PowerTransformer to revise a set of modern movie scripts and successfully mitigate the bias we previously uncovered [9], raising the power and agency that female characters are portrayed with.

Future work My work has taken steps towards representing and mitigating social biases implied or stated in text, but much is left to be done. I am particularly interested in designing new controllable debiasing models that make use of more structured representations of social biases (e.g., Social Bias Frames). This also requires designing new methods for evaluating these debiasing system, e.g., using machine-in-the-loop writing setups. Additionally, I want to continue investigating how we can improve neural models to better handle structured inferences like in Social Bias Frames [8], and particularly ones that can handle unstated implications instead of relying on lexical cues. Finally, I want to create more holistic formalisms of social biases that take a statement’s context into account through categorical and free-text variables. For example, not only can preceding sentences change a statement’s interpretation, but the social context (e.g., speaker identity) also influences pragmatic and biased implications [11].

3 Diagnosing the Ethics and Fairness of NLP systems

Many modern AI systems exhibit biases that prevent fair and safe deployment. One important part of my research is quantitative and qualitative analysis of AI systems' fairness to ensure that they are not exacerbating existing unbalanced power structures and harming minority populations. This requires investigating the behavior of these systems, as well as the training data, procedure, and learning objective.

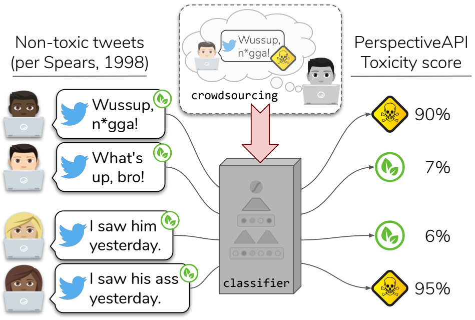

Figure 4: An illustration of how hate speech detection can discriminate against racial minorities by flagging harmless speech as toxic. Examples from [24].

In my ACL 2019 work [11], I uncovered a **strong racial bias in hate speech detection systems**, despite their aim to protect the very minorities they are censoring. We showed that hate speech classifiers consistently flag speech by African Americans as toxic compared to other races (Figure 4), as measured by widely used ML fairness criteria [25]. After tracing the origin of these biases to the dataset collection, we proposed a possible solution by **creating an annotation framework that reduces racial biases** by highlighting the dialect or race of a tweet's author. This work was nominated for the ACL 2019 best short paper award.

In follow up work [13], we adapted several data and model debiasing methods (e.g., filtering, ensemble objectives) to the toxicity detection task to assess their usefulness for mitigating racial biases. Unfortunately, our findings showed that these methods have limited effectiveness, confirming that dataset quality is crucial to fair toxicity detection systems.

My research has also investigated biases in general purpose NLP systems. Recently [12], colleagues and I **examined the risk of toxic or biased language *de*generation by widely used pretrained language models** [e.g., GPT-2, GPT-3; 26, 27]. To study this phenomenon, we introduced RealToxicityPrompts, a test bed of 150k natural language prompts that could induce toxicity. Using these prompts, we showed that current models risk devolving into generating toxic, hateful, rude, or profane text, even when using state-of-the-art controllable generation methods adapted to avoid toxicity. We traced toxic degeneration back to models' training data, for example, finding non-trivial amounts of fake news articles and toxic documents in GPT-2's [26] training data.

Future work Looking forward, I plan to continue designing methods to diagnose and improve the cultural awareness and fairness of NLP systems using human-centered approaches [28, 29], to avoid backfiring against minority groups. For example, I want to explore ways to make culturally aware toxicity detection systems, which requires working with social scientists to understand how user factors (e.g., race, politics) influence the understanding of toxicity. Additionally, I want to design more holistic and ethical learning and data collection paradigms that incorporate social variables and user factors. For example, I am actively investigating the training dynamics of toxicity in language models, to quantify the causal link between toxicity in training data and in generated text.

References

[1] **Maarten Sap**, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[2] **Maarten Sap***, Hannah Rashkin*, Derek Chen, Ronan LeBras, and Yejin Choi. SOCIAL IQA: commonsense reasoning about social interactions. In *EMNLP*, 2019.

[3] Antoine Bosselut, Hannah Rashkin, **Maarten Sap**, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, 2019.

[4] Hannah Rashkin*, **Maarten Sap***, Emily Allaway, Noah A. Smith, and Yejin Choi. EVENT2MIND: Commonsense inference on events, intents, and reactions. In *ACL*, 2018.

[5] Hannah Rashkin, Antoine Bosselut, **Maarten Sap**, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *ACL*, 2018.

[6] Hao Fang, Hao Cheng, **Maarten Sap**, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. Sounding board: A user-centric and content-driven social chatbot. In *NAACL System Demonstrations*, 2018.

[7] **Maarten Sap**, Eric Horvitz, Yejin Choi, Noah A Smith, and James W Pennebaker. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *ACL*, 2020.

[8] **Maarten Sap**, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. SOCIAL BIAS FRAMES: Reasoning about social and power implications of language. In *ACL*, 2020.

[9] **Maarten Sap**, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *EMNLP*, 2017.

[10] Xinyao Ma*, **Maarten Sap***, Hannah Rashkin, and Yejin Choi. POWERTRANSFORMER: Unsupervised controllable revision for biased language correction. In *EMNLP*, 2020.

[11] **Maarten Sap**, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *ACL*, 2019.

[12] Samuel Gehman, Suchin Gururangan, **Maarten Sap**, Yejin Choi, and Noah A. Smith. REALTOXICITYPROMPTS: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, 2020.

[13] Xuhui Zhou, **Maarten Sap**, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. The ineffectiveness of algorithmic debiasing for toxic language detection. In *in submission at EACL*, 2021.

[14] David Gunning. Machine common sense concept paper. October 2018.

[15] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA, 2013. ACM.

[16] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103, 2015.

* denotes equal contribution between first two authors.

[17] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. In *AAAI*, 2020.

[18] Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Predictive coding approximates backprop along arbitrary computation graphs. June 2020.

[19] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. July 2020.

[20] Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. Contextual affective analysis: a case study of people portrayals in online #MeToo stories. In *ICWSM*, 2019.

[21] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. A framework for the computational linguistic analysis of dehumanization. *arXiv preprint arXiv:2003.03014*, 2020.

[22] Maria Antoniak, David Mimno, and Karen Levy. Narrative paths and negotiation of power in birth stories. In *CSCW*, 2019.

[23] Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312, 2020.

[24] Arthur K Spears. African-American language use: Ideology and so-called obscenity. In Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh, editors, *African-American English: Structure, History and Use*, pages 226–250. Routledge New York, 1998.

[25] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[27] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are Few-Shot learners. May 2020.

[28] Alexa Hagerty and Igor Rubinov. Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. July 2019.

[29] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. Human-Centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA, April 2020. Association for Computing Machinery.