Comparing Humans and Models on a Similar Scale: Towards Cognitive Gender Bias Evaluation in Coreference Resolution

Gili Lior (gili.lior@mail.huji.ac.il)
School of Computer Science and Engineering,
The Hebrew University of Jerusalem, Israel

Gabriel Stanovsky (gabriel.stanovsky@mail.huji.ac.il)
School of Computer Science and Engineering,
The Hebrew University of Jerusalem, Israel

Abstract

Spurious correlations were found to be an important factor explaining model performance in various NLP tasks (e.g., gender or racial artifacts), often considered to be "shortcuts" to the actual task. However, humans tend to similarly make quick (and sometimes wrong) predictions based on societal and cognitive presuppositions. In this work we address the question: *can we quantify the extent to which model biases reflect human behaviour?* Answering this question will help shed light on model performance and provide meaningful comparisons against humans. We approach this question through the lens of the *dual-process theory* for human decision-making. This theory differentiates between an automatic unconscious (and sometimes biased) "fast system" and a "slow system", which when triggered may revisit earlier automatic reactions. We make several observations from two crowdsourcing experiments of gender bias in coreference resolution, using self-paced reading to study the "fast" system, and question answering to study the "slow" system under a constrained time setting. On real-world data humans make ∼3% more gender-biased decisions compared to models, while on synthetic data models are ∼12% more biased. We make all our of our code and data publicly available.[1]

Introduction

Recent work has identified that large-scale models achieve impressive results on various natural language benchmarks by exploiting correlations which do not seem semantically meaningful for solving the task (Gururangan et al., 2018; Gardner et al., 2021). Leveraging such *spurious correlations* is often considered an indication that models do not solve the actual task, but instead resort to finding statistical "shortcuts" around the problem (Geva et al., 2021; Savoldi, Gaido, Bentivogli, Negri, & Turchi, 2021).

In parallel, works in cognitive psychology identify that finding shortcuts may in fact be a *feature* of human intelligence, which on the one hand helps us cope with missing or implicit information (H. P. Grice, 1975; P. Grice, 1989), while on the other hand may also lead to harmful behavior. In the context of gender bias in coreference resolution, which will be the focus of our work, studies have found that human subjects tend to prefer the stereotypical reading in various modalities, such as event-related brain potentials, reading times, or eye movements (Osterhout, Bersick, & McLaughlin, 1997; Kennison & Trofe, 2003; Duffy & Keir, 2004).

In this work we propose to integrate findings from these two lines of research and quantify the extent to which model biases resemble human behavior. We distinguish between two ends of a spectrum. On the one hand we place *annotation artifacts*, which hold only in specific training sets, e.g., associating the word "cat" with contradiction in NLI (Gururangan et al., 2018). On the other hand of the spectrum we place *human-like biases* which are sometimes useful in real-world scenarios (e.g., in common sense reasoning (Lent & Søgaard, 2021)), but also produce harmful, unwanted behavior (as in gender bias (Schwartz & Stanovsky, 2022)). These are likely to arise in any real-world dataset, and may require subtle debiasing techniques in either modelling or data collection.

To place model biases on this spectrum, we develop human annotation interfaces and derive evaluation metrics which compare between humans and models, thus putting them on the same scale. In particular, we focus on gender bias in coreference resolution in the English language, which was widely studied in machine learning and psycholinguistics, allowing us to explore results in the intersection of these areas.

To achieve this, we study human biases through the lens of the *dual-process theory* (Evans, 2008), which posits that there are two cognitive systems participating in humans' decision making process. *System 1* is fast, associative and automatic, while *System 2* is slow, conscious and effortful. *System 1* heuristics are considered a survival mechanism. Humans make thousands of decisions a day, and if all of them were consciously processed, our brain would not handle the cognitive load. But on the other hand, when *System 1* "shortcuts" are wrong and *System 2* does not revise it, erroneous and biased decisions may occur (Kahneman, 2011).

Within this framework, we propose two human experiments to quantify the heuristics made by *System 1*. The first experiment tests *System 1* directly, by examining how gender bias manifests in self-paced reading (Jegerski, 2013), which approximate eye tracking, largely considered to be an unconscious process (Rayner, 1998). The second experiment is question answering (QA) over coreference-related questions. QA is likely to invoke *System 2*, as it requires more conscious effort (Wang & Gafurov, 2003). We then add different artificial time constraints, to examine how *System 1* heuristics are expressed in a task that requires more cognitive effort.

Finally, we crowdsource annotations for the two experiments over synthetic and real-world sentences, and make several important observations, comparing humans to two state-of-the-art coreference models. Both experiments sur-

[1] https://github.com/SLAB-NLP/Cog-GB-Eval

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

face comparable gender biases to those shown by models. Specifically, in the QA experiment over the natural sentences, models' overall accuracy is significantly lower than humans, but both show similar biases. In contrast, for the synthetic sentences, the models' overall accuracy was closer to humans, but models have shown larger gender bias.

To the best of our knowledge our work presents a first quantitative evaluation of gender bias in coreference resolution models versus human behavior, specifying the conditions needed to elicit comparable biases from humans through time constraints. Our results indicate that model biases indeed resemble decisions made by humans with restricted attention span. Future work may leverage our evaluation paradigm and revisit it for other tasks and future models.

Background

We begin by describing previously published datasets designed to test model biases in coreference resolution . To measure human performance, we then discuss Maze (Forster, Guerrera, & Elliot, 2009), a self-paced reading approach approximating eye-tracking measurements.

Gender Bias Datasets

We use three coreference gender bias datasets as outlined below, and summarized in Table 1.

WinoBias (Zhao, Wang, Yatskar, Ordonez, & Chang, 2018) and Winogender (Rudinger, Naradowsky, Leonard, & Van Durme, 2018) consist of 3,888 synthetic, short sentences. Each of the sentences conforms to a similar template consisting of two entities, identified by their profession, and a single referring pronoun. The datasets are balanced with respect to stereotypical gender-role assignment (e.g., female secretaries) versus non-stereotypical assignment (e.g., male nurses). These datasets are good for controlled experiments but consist of a small variety of linguistic constructions, and do not represent real-world distributions.

In contrast, the BUG corpus (Levy, Lazar, & Stanovsky, 2021) aims to find such templates "in the wild". It consists of 1,720 sentences sampled from natural corpora (e.g., Wikipedia and PubMed) and better approximates real-world distribution in terms of sentence length, vocabulary and gender-role stereotypes. Similar to Winogender and WinoBias, each sentence in BUG presents entities identified via their profession and a referring pronoun. BUG also provides a binary annotation for each sentence marking whether is conforms to societal norms. For accuracy sake, we use a subset of BUG which was manually annotated.

Maze

For our proposed evaluation metric presented in the Experiments section, we use Maze (Forster et al., 2009), a platform for measuring self-paced reading (Jegerski, 2013). This platform is an alternative for eye-tracking measurements (Witzel, Witzel, & Forster, 2012), that does not require specialized equipment and in-house annotators. Instead, Maze can be easily deployed on crowdsourcing platforms, allowing us to collect annotations at scale.

As exemplified in Figure 2, Maze iteratively presents two options for the next word in a sentence, and a human annotator needs to select the most probable alternative given previously seen words. The time for choosing the correct word approximates its reading time.

	Original		QA		MAZE	
	#pro	#anti	#pro	#anti	#pro	#anti
WinoBias	1582	1586	756	717	607	603
Winogender	216	216	203	216	35	35
BUG	865	420	431	271	565	315

Table 1: Statistics for coreference gender bias datasets. "Original" presents the number of sentences in each of the datasets. "QA" and "MAZE" show the number of sentences in our experiments, further decomposed into pro-stereotypical and anti-stereotypical sentences. The reduction in sampling sizes is due to additional filtering and distribution tuning. See the Experiments section for more details.

Working Definitions

In this section, we formally define key concepts commonly used throughout the paper.

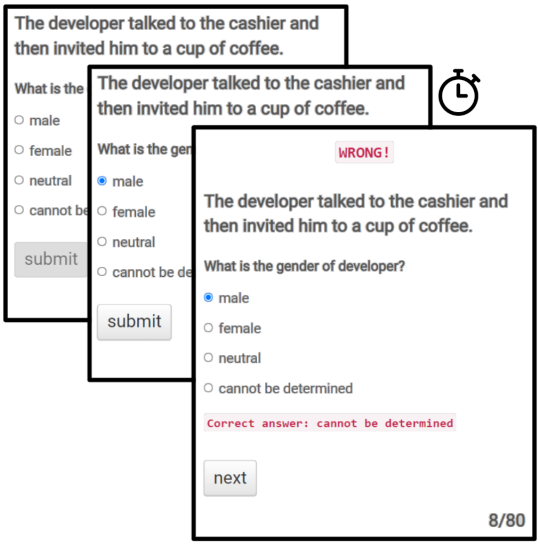
Gender. We use existing gender bias corpora, as described in the Background section, using pronouns with three grammatical genders: feminine, masculine, and neutral. These datasets are generally devoid of other types of pronouns, such as neo-pronouns.[2] Collecting corpora for diverse types of pronouns is left as an important avenue for future work, e.g., as outlined by (Lauscher, Crowley, & Hovy, 2022).

Pro-stereotype/Anti-stereotype. A coreference relation between a pronoun and an entity in a sentence is deemed pro-stereotypical if the referring pronoun's gender conforms to societal norms (e.g., "nurse" and "she"), otherwise it is marked anti-stereotypical (e.g., "cleaner" and "he"). These definitions naturally extend to sentences with a single pronoun. These are deemed pro or anti stereotypical according to the relation between the entity and its referring pronoun. To estimate the stereotypical gender norm per profession we use labels provided in the previously-published gender bias datasets (Rudinger et al., 2018; Zhao et al., 2018), based on both human annotations and reports published by The U.S. Bureau of Labor Statistics.[3]

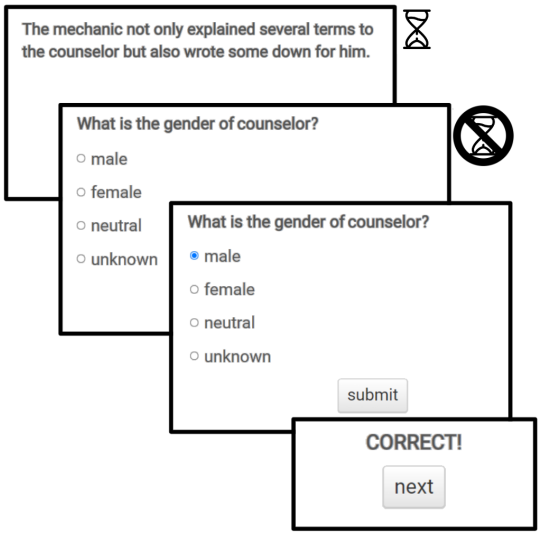
Gender bias. We adopt the *Historical Bias* definition (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021): *"the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection"*. This definition connects between the physical world and how

[2] https://www.unf.edu/lgbtqcenter/Pronouns.aspx

[3] We note that these norms may vary between cultures. Adapting them to other communities is left for an important future work.



(a) QA calibration interface. A sentence and a question are shown for an unlimited time. After submitting the answer, a feedback is shown on screen, including the correct answer. We record the participants choice and the time they took to answer.



(b) QA Experiment interface. A sentence is shown for a limited time. Then, the sentence disappears and only a question is shown, for an unlimited time. After submitting an answer, a feedback message is shown. Here we only record the the participants choice.

Figure 1: QA calibration and main experiment interfaces.

it manifests in the the training data. In particular, historical bias appears when models make predictions based on the gender distribution in the training data, rather than the relations between entities in the sentence.

Experiments

In the following section we present our two experiments for measuring human biases. The design choices we make follow common practices in psycholinguistic literature.

QA Experiment

In this experiment we present a sentence followed by a multiple-choice question regarding the gender of an entity in the sentence, eliciting coreference resolution decisions. For example, given the sentence "The developer talked to the cashier and invited him for a cup of coffee", and the question: "What is the gender of the cashier?", the four possible answers are 'male', 'female', 'neutral' and 'unknown', and the expected answer is 'male'.

QA is likely to invoke *System 2* as it involves conscious decision making. To test *System 2* under a constrained setting, annotators observe the sentence for a limited time before it disappears and then they can answer the question. See Figure 1 for an example annotation interface.

Filler questions. Following common practice in human annotation tasks, we introduce filler questions to prevent participants from focusing on certain aspects of the sentence (e.g., its pronoun). We automatically formulate questions on predicate-argument relations using a pretrained QA-SRL model (FitzGerald, Michael, He, & Zettlemoyer, 2018), that produces different question formats, e.g., asking about the subject ("who might be talking?"), object ("who was being hired?") and other entities in the sentence. Similarly to other psycholinguistics works, our filler questions constitute 50% of the total questions in the experiment (Witzel et al., 2012; Kim, Gabriel, & Gygax, 2019; Boyce, Futrell, & Levy, 2020).

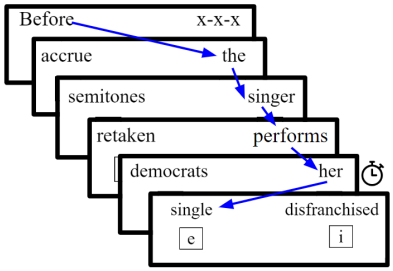


Figure 2: MAZE experiment interface. At each step, participants need to distinguish the next word from a distractor, by pressing the correct keyboard key. We record the time they took for identifying the pronoun.

Calibration. To account for different reading paces (Rayner, Schotter, Masson, Potter, & Treiman, 2016), we begin with calibrating a baseline reading pace for each participant. We present the sentence for an unlimited time along with the question and measure the time it takes the participant to submit the correct answer. This is then normalized by the sentence length (in words) to approximate a participant's reading pace. See Figure 1a for an example of this interface.

Annotation interface. Each participant observes a single sentence for a limited amount of time. Then, the sentence

		Wino			BUG		
		pro	anti	Δ_{QA}	pro	anti	Δ_{QA}
Humans baseline	0.75	89.8	89.7	**0.1**	87.8	82.8	**5.0**
	0.50	88.8	87.6	**1.2**	85.7	79.1	**6.5**
	0.25	78.7	78.5	**0.2**	82.6	75.7	**6.9**
Models	SpanBERT	86.5	73.4	**13.1**	62.2	60.0	**2.2**
	s2e	91.3	77.7	**13.6**	61.7	59.3	**2.4**

Table 2: Human and model results in the QA task on the same sentences. 'pro' and 'anti' columns show results on pro-stereotypical and anti-stereotypical gender questions. Δ_{QA} stands for the difference between the two categories (pro minus anti), indicating biased performance, approximating *System 2* biases.

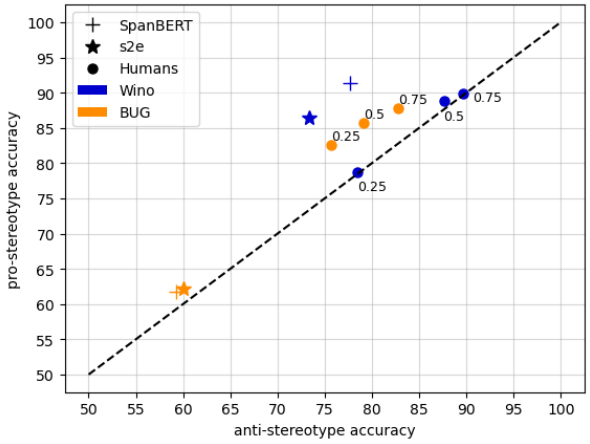


Figure 3: Visualization for the results shown in Table 2. The x-axis is the performance over the anti-stereotypical sentences, and the y-axis is the performance over the pro-stereotypical sentences. Values above the dashed black line show gender biased performance. Datasets are represented by color, while humans are distinguished from models by the indicator's of your shape. $\{0.25, 0.5, 0.75\}$ are the fractions of the baseline reading pace given to humans. All evaluations found some degree of gender bias.

disappears and a question regarding one of the entities in the sentence is shown for an unlimited time. The time each sentence is presented on screen is calculated by $(\alpha \cdot avg \cdot l)$, where *avg* is the participant's reading pace, *l* is the length of the sentence in words, and α is sampled i.i.d from $\{0.25, 0.5, 0.75\}$ to present the sentence for a fraction of the participant's pace. An example of this interface is shown in Figure 1b.

Annotator feedback. Following (Malmaud, Levy, & Berzak, 2020), we show participants a feedback message indicating if they were correct after every submitted answer, both for filler questions and for the actual task. Feedback in multiple-choice questions has been shown to improve performance and reduce low-quality annotations (Butler & Roediger, 2008). To mitigate the risk of affecting responses in unintended ways, we use filler questions that prevent annotators from overspecializing in the task.

Filtering non-coreference errors. This setup may produce errors which do not relate to coreference. For example, answering that the gender of the entity is masculine while the presented pronoun is feminine (and vice versa) does not indicate a coreference error, and is therefore ignored.

Self-Paced Reading Experiment

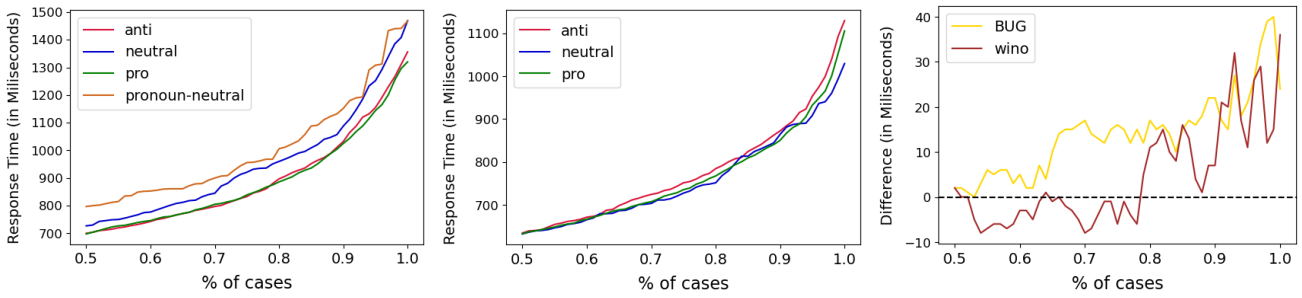
In the second experiment, we approximate trends in reading time of pronouns in pro-stereotypical versus anti-stereotypical instances, which is considered an unconscious process, and hence a good proxy for *System 1*'s biases (Rayner, 2009).

We use MAZE to approximate the time it takes a participant to choose the pronoun in our sentences (see Figure 2). This implicitly measures the timing of a coreference decision since the pronoun indicates the gender of a previously mentioned entity. Previous work has identified that self-paced reading is a good proxy for natural reading when comparing between readings of different sentences, albeit it may overestimate the absolute reading times (Yan & Jaeger, 2020). This makes self-paced reading adequate for our purposes, as we are interested in the *trends* shown in response time between pro-stereotypical and anti-stereotypical instances.

Filtering ambiguous instances. Since MAZE presents the words in a linear order, we note that there are instances when the pronoun appears before the context needed to infer its antecedent. E.g., when reading the prefix "The sheriff questioned the housekeeper as *she*..." it is yet unclear whether "she" refers to the sheriff (e.g., as in "...*she* needed to find the thief.") or the housekeeper (e.g., in "... *she* was cleaning"). Since the reader cannot know which of the suffixes will follow, these instances do not reflect gender bias decisions. To address this issue in WinoBias and Winogender, we sample only sentences where the pronoun appears after all verbs in the sentence, e.g., "The tailor thought the janitor could be good at sewing and encouraged her". In a preliminary analysis we find that this heuristic may be over-strict, but leads to high precision, which was most important for our analyses. From BUG we sampled only sentences where the pronoun appeared after its antecedent. We find that this sampling works well for the sentences in BUG, which usually consist of a single entity.

For the synthetic sentences, this sampling produces a subset of 1,335 viable sentences. Most of the instances which were filtered out come from Winogender, because in most of its sentences the pronoun appears before one of the verbs in the sentence. For BUG, this sampling produces a subset of 1,603 viable sentences.

Annotation interface. At each time step participants are shown two possible words, and they need to choose the next word in the sentence according to previous context. See Figure 2 an example. We allow participants to retry in case of an error, and record the time until their first answer, as well as the total time until the correct option was chosen.



(a) Wino (b) BUG (c) delta-stereotype (anti minus pro)

Figure 4: Figures 4a and 4b show the CDF of response times needed to distinguish the pronoun from its distractor in MAZE. I.e., coordinate (x, y) on the graph implies that x% of the annotations required a response time of y ms or less. Figure 4c shows Δ_{MAZE} for humans, i.e., the difference between anti-stereotypical response time and pro-stereotypical response time, where values above $y = 0$ indicate gender biased performance.

Simulating arbitrary time limitations. Similarly to the QA experiment, we would like to introduce a notion of time constraints. If we would have limited the amount of time given to distinguish the next word, a participant would either: (a) choose correctly (b) choose incorrectly (c) not respond in time. Instead of testing participants over different discrete time limitations, we make the following assumption: if a participant's response time for a correct annotation was x ms, any time limit below x ms would not be enough time for responding (option (c) above). Following, we do not limit participants reading time, but instead compute a cumulative distribution function over all possible observed response times. Finally, we use A-Maze (Boyce et al., 2020) to automatically generate probable distractors.

Results

In this section we summarize the main findings from our two crowdsourcing experiments. We find that the overall human accuracy for both tasks was good, reaching 94.48% on the gender questions in unrestricted QA, and 98.13% in MAZE, indicating an understandable task and high quality annotations.

Experimental Setup

The two experiments collected annotations from 33 participants on the Amazon Mechanical Turk platform. Our average hourly pay was 8.53 USD. The overall cost to produce our annotations was 1,030 USD. To qualify, workers had to have at least 5,000 accepted HITs at an acceptance rate of at least 96%, and hail from English-speaking countries. In addition, we ran a qualification HIT which required workers to score at least 85% on an unconstrained version of the QA task. Following (von der Malsburg, Poppels, & Levy, 2020), we annotated $3K$ instances with gender bias signal for each experiment and each dataset, amounting to $12K$ annotations. We deploy the QA task using Anvil,[4] and the MAZE task using Ibex.[5] Finally, we use the IQR technique to remove outliers in the self-paced reading (Vinutha, Poornima, & Sagar, 2018), which may arise due to network connectivity issues.

Evaluation Metrics

Following previous work, we compute gender bias as the difference in performance between pro-stereotypical and anti-stereotypical instances (Stanovsky, Smith, & Zettlemoyer, 2019). In the QA task we denote as Δ_{QA} the difference between accuracy on pro-stereotypical versus anti-stereotypical gender questions, which is a proxy for constrained *System 2* gender bias. In the self-paced reading task we compute the difference in response time to identify the pronoun, marked as Δ_{MAZE}, and is a proxy for *System 1* biases. For consistency, both metrics are defined such that larger values indicate more gender biased performance. I.e., for Δ_{MAZE} we subtract the response time for pro-stereotypical instances from the anti-stereotypical instances, as longer response times indicate worse performance.

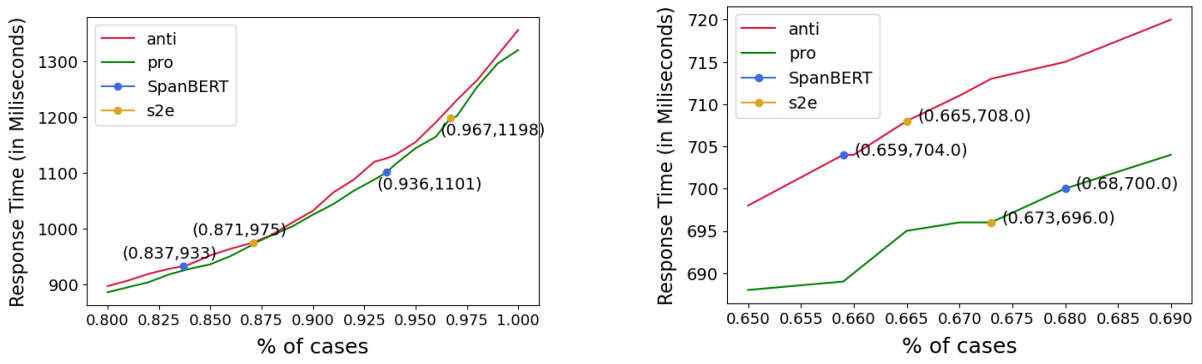
QA results

Several observations can be drawn from the results for the QA task, presented in Table 2 and visualized in Figure 3, showing the biases caused by limiting the resources of *System 2*.

Human subjects show more gender bias as they are given less time to read the sentence. For both natural and synthetic sentences, we find that Δ_{QA} for humans increases between when they are given 0.75 and 0.5 of their baseline reading pace, and for natural sentences specifically we see this increase also between 0.5 and 0.25. I.e., the difference in performance between pro-stereotypical and anti-stereotypical increases the less time participants have. However at some point, participants will not have enough time to process the sentence. This is observed in Winogender and WinoBias when $\alpha = 0.25$, where human performance equally degrades across both anti-stereotypical and pro-stereotypical, in parallel with an increase in non-coreference errors from around

[4] https://anvil.works/

[5] https://github.com/addrummond/ibex



(a) Wino

(b) BUG

Figure 5: Model performance versus human annotations. The blue and yellow points are the intersection points with the different models' accuracy and their matching category threshold. For example, the blue point intersecting the red line, is the human threshold that matches SpanBERT accuracy on anti-stereotypical sentences.

2% when $\alpha \in [0.5, 0.75]$ to 5% when $\alpha = 0.25$.

Human subjects were found more prone to gender biased answers on naturally-occurring sentences. Table 2 shows larger Δ_{QA} for natural sentences than for the synthetic ones, and in Figure 3 the points representing human performance on BUG are farther from the diagonal, indicating more biased performance. This may stem from the templated nature of the synthetic sentences which allows subjects to master them.

Self-Paced Reading Results

Several conclusions are drawn from this results of this experiment, shown in Figure 4, approximating *System 1* biases.

Higher human gender bias is observed the more processing time is needed. Figures 4a and 4b show the CDF of response times for distinguishing the pronoun from a distractor over the correct annotations (which consist of 98% of all annotations). Figure 4c shows Δ_{MAZE}, which is the difference between anti-stereotypical and pro-stereotypical instances in 4a and 4b. The longer the response time allowed (and hence more annotations are counted), a more pronounced Δ_{MAZE} is observed.

Human gender bias was observed only when accounting for at least 80% of the synthetic sentences. Positive Δ_{MAZE} indicates longer response time for anti-stereotypical sentences than pro-stereotypical, and so considered gender bias. Figure 4c shows that Δ_{MAZE} is positive after accumulating 80% of the annotations on the CDF curve, while for the natural sentences this effect is found after 50% of annotations.

Comparing Model and Human Biases: Discussion and Conclusions

We evaluate SpanBERT (Joshi et al., 2020) and *s2e* (Kirstain, Ram, & Levy, 2021) on the same sentences annotated by humans in each of the tasks, and compare the bias in results between humans and models. Below we outline several key findings.

Models exhibit gender bias more than humans on synthetic sentences in the QA experiment. Table 2 shows that Δ_{QA} on Winogender and WinoBias is larger for models when compared to humans on any fraction of the reading pace. Additionally, Figure 3 shows that over Winogender and WinoBias, models are farther from equilibrium line than humans, and the human performance on anti-stereotypical instances is superior to models. This may indicate that to achieve good performance, models rely on gender bias more than humans.

Models show more gender bias on synthetic sentences than in real-world sentences, as opposed to humans where gender bias is more pronounced over natural sentences. For the QA experiment this trend is seen in Δ_{QA} columns in Table 2. As for the self-paced reading task, Figure 4c shows that human's Δ_{MAZE} on BUG is above Δ_{MAZE} on Winogender and WinoBias, while for models Δ_{MAZE} is the distance on the x-axis between the points in Figures 5a and 5b, which is smaller on BUG for both models. In humans, this may arise due to mastering synthetic sentences to the point they do not rely on gender stereotypes to excel in it. In contrast, the degraded performance of models on real-world sentences diminishes the gains from biased predictions.

Conclusion: Model biases reflect human decision-making under constrained settings. Revisiting our research question, our findings suggest that gender bias in coreference resolution is comparable to human biases rather than an annotation artifact, indicating it will likely creep up in real-world datasets along with other, more desired human behavior, like common sense reasoning.

Future Work. Follow-up work may compare our results with competing cognitive theories, e.g., (Bursell & Olsson, 2021), as well as developing some kind of "slow reasoning" models, e.g., via early exiting (Schwartz, Stanovsky, Swayamdipta, Dodge, & Smith, 2020; Laskaridis, Kouris, & Lane, 2021).

Acknowledgments

We thank Yevgeni Berzak and Roger Levy for many fruitful discussions, and the anonymous reviewers for their valuable feedback. This work was supported in part by a research gift from the Allen Institute for AI, and a research grant 2336 from the Israeli Ministry of Science and Technology.

References

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Bursell, M., & Olsson, F. (2021). Do we need dual-process theory to understand implicit bias? a study of the nature of implicit bias against muslims. *Poetics*, *87*, 101549.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, *36*(3), 604–616.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

FitzGerald, N., Michael, J., He, L., & Zettlemoyer, L. (2018, July). Large-scale QA-SRL parsing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2051–2060). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1191 doi: 10.18653/v1/P18-1191

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, *41*(1), 163–171.

Gardner, M., Merrill, W., Dodge, J., Peters, M., Ross, A., Singh, S., & Smith, N. A. (2021, November). Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1801–1813). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.135 doi: 10.18653/v1/2021.emnlp-main.135

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, *9*, 346–361. Retrieved from https://aclanthology.org/2021.tacl-1.21 doi: 10.1162/tacl_a_00370

Grice, H. P. (1975). Logic and conversation. In *Speech acts.*

Grice, P. (1989). *Studies in the way of words*.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018, June). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 107–112). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2017 doi: 10.18653/v1/N18-2017

Jegerski, J. (2013). Self-paced reading. In *Research methods in second language psycholinguistics* (pp. 36–65). Routledge.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77. Retrieved from https://aclanthology.org/2020.tacl-1.5 doi: 10.1162/tacl_a_00300

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, *32*(3), 355–378.

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the internet-based instrument psytoolkit: A comparison between web-based (psytoolkit) and lab-based (e-prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PloS one*, *14*(9), e0221802.

Kirstain, Y., Ram, O., & Levy, O. (2021, August). Coreference resolution without span representations. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 14–19). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-short.3 doi: 10.18653/v1/2021.acl-short.3

Laskaridis, S., Kouris, A., & Lane, N. D. (2021). Adaptive inference through early-exit networks: Design, challenges and directions. *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*.

Lauscher, A., Crowley, A., & Hovy, D. (2022). Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv:2202.11923*.

Lent, H., & Søgaard, A. (2021, November). Common sense bias in semantic role labeling. In *Proceedings of the seventh workshop on noisy user-generated text (w-nut 2021)* (pp. 114–119). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wnut-1.14 doi: 10.18653/v1/2021.wnut-1.14

Levy, S., Lazar, K., & Stanovsky, G. (2021, November). Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 2470–

2480). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.211 doi: 10.18653/v1/2021.findings-emnlp.211

Malmaud, J., Levy, R., & Berzak, Y. (2020, November). Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th conference on computational natural language learning* (pp. 142–152). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.conll-1.11 doi: 10.18653/v1/2020.conll-1.11

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of eye movement research*, *2*(5), 1.

Rayner, K., Schotter, E. R., Masson, M. E. J., Potter, M. C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, *17*(1), 4-34. (PMID: 26769745) doi: 10.1177/1529100615623267

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018, June). Gender bias in coreference resolution. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 8–14). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2002 doi: 10.18653/v1/N18-2002

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. Retrieved from https://aclanthology.org/2021.tacl-1.51 doi: 10.1162/tacl_a_00401

Schwartz, R., & Stanovsky, G. (2022, July). On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the association for computational linguistics: Naacl 2022* (pp. 2182–2194). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-naacl.168 doi: 10.18653/v1/2022.findings-naacl.168

Schwartz, R., Stanovsky, G., Swayamdipta, S., Dodge, J., & Smith, N. A. (2020). The right tool for the job: Matching model and instance complexities. In *Annual meeting of the association for computational linguistics*.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019, July). Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1679–1684). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1164 doi: 10.18653/v1/P19-1164

Vinutha, H., Poornima, B., & Sagar, B. (2018). Detection of outliers using interquartile range technique from intrusion dataset. In *Information and decision sciences* (pp. 511–518). Springer.

von der Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 united states and 2017 united kingdom elections. *Psychological science*, *31*(2), 115–128.

Wang, Y., & Gafurov, D. (2003). The cognitive process of comprehension. In *The second ieee international conference on cognitive informatics, 2003. proceedings.* (pp. 93–97).

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of psycholinguistic research*, *41*(2), 105–128.

Yan, S., & Jaeger, T. F. (2020). Expectation adaptation during natural reading. *Language, Cognition and Neuroscience*, *35*(10), 1394–1422.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018, June). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 15–20). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2003 doi: 10.18653/v1/N18-2003